Transactions

of the I·R·E

Professional Group on

INFORMATION THEORY

Report of Proceedings

SYMPOSIUM ON INFORMATION THEORY

London, England

September, 1950

Published by Special Arrangement
with the Ministry of Supply
London

Price per copy—

Members of the IRE Professional Group
on Information Theory—$2.25
Members of the IRE—$3.40
Nonmembers—$6.75

Lithoprinted in the United States of America

PGIT—1

FEBRUARY, 1953

The Institute of Radio Engineers

SYMPOSIUM ON INFORMATION THEORY

SEPTEMBER 1950

(Held in the Lecture Theatre of the Royal Society
Burlington House, by kind permission of
the President and Council.)

REPORT OF PROCEEDINGS

PUBLISHED BY SPECIAL ARRANGEMENT
WITH THE MINISTRY OF SUPPLY.
LONDON.

PREFACE

This is the first in a series of publications which it is planned to make available to the members of the Professional Group on Information Theory of the IRE.

The delay in distributing our first publication was brought about partially in the interests of keeping the costs to a minimum, and the somewhat unusual problems in our handling, as our first offering, a publication originating from abroad. The non-standard size could not readily be avoided without increased costs.

It is believed that subsequent publications will be provided more promptly.

Appreciation should be expressed to Melpar, Inc. of Alexandria, Virginia, and Federal Telecommunication Laboratories, Inc. of Nutley, New Jersey, for their assistance and cooperation which made the reproduction possible within the budget amount.

We wish in particular to thank Mr. F. S. Barton, Principal Director of Electronics Research and Development, Ministry of Supply, and Prof. Willis Jackson of the Imperial College of Science and Technology for their assistance in obtaining authorization for this publication.

Acknowledgements should also be tendered to Dr. W. G. Tuller of Melpar, Mr. N. Marchand of Marchand Electronics Laboratories, and Mr. J. Rubino of Federal Telecommunication Laboratories, for their part in this project.

L. A. deRosa
Chairman,
Publications Committee
Professional Group, Information Theory

FOREWORD

During the past hundred years or thereabouts a variety of techniques have been devised for transmitting messages electrically from point to point. It is only of recent years, however, through the work of a few theoretically minded communication engineers that means have been provided for assessing quantitatively the commodity which is transmitted, namely, the "information" content of messages, and of determining the extent to which existing techniques fall short of what may be attainable. This recent work proves to have a significance well outside the sphere of Electrical Communications. A new branch of science is emerging which reveals and clarifies connections between previously largely unrelated fields of research concerned with different aspects of the processes by which living organisms - in particular man - collect, classify, convert and transmit information. A confluence of different fields of investigation is of course no new phenomenon in the history of science, but the wide recognition of its occurrence can seldom have been so rapid as in the present case.

This Symposium was organised to afford an opportunity for discussion of the nature and potentialities of this recent work among interested scientists and engineers. Those who participated comprise groups of mathematicians, statisticians, physicists, biologists, physiologists, psychologists and electrical engineers drawn from this country and abroad, and all will wish me to express gratitude to Dr. Claude E. Shannon of the Bell Telephone Laboratories, New Jersey and Prof. J.B. Wiesner of the Electronics Laboratory, Massachusetts Institute of Technology for their stimulating contributions.

I take this opportunity of thanking also the distinguished men who acted as Chairmen of the various sessions, the Ministry of Supply and the British Broadcasting Corporation for their generous contributions to the expenses of the Symposium, and the Royal Society for its encouragement and for permitting the use of its lecture theatre.

Willis Jackson.

Electrical Engineering Department,
Imperial College,
London, S.W.7.
September, 1950.

INDEX

PROGRAMME

Tuesday, 26th September, 1950

Chairman - Professor Sir David Brunt, Sec. R.S.

1)	A History of the Theory of Information	E. C. Cherry, Electrical Engineering Dept., Imperial College, London.
2)	Communication Theory - Exposition of Fundamentals	Dr. C.E. Shannon, Bell Telephone Labs., New Jersey, U.S.A.

Chairman - Professor H. M. Massey, F.R.S.

3)	Communication Theory and Physics	Dr. D. Gabor, Electrical Engineering Dept., Imperial College, London.
4)	Quantal Aspects of Scientific Information	D. M. MacKay, King's College, London.

Wednesday, 27th September, 1950

Chairman - Professor R. A. Fisher, F.R.S.

5)	The Statistical Approach to the Analysis of Time Series	Professor M. S. Bartlett, Manchester University.
6)	General Treatment of the Problem of Coding. The Lattice Theory of Information.	Dr. C. Shannon, Bell Telephone Labs., New Jersey, U.S.A.

Chairman - Dr. R. Cookburn.

7)	Theory of Radar Information	P. M. Woodward, T.R.E., Malvern.
8)	Fluctuations and Theory of Noise	Dr. D. K. C. MacDonald, Clarendon Laboratory, Oxford.

Thursday, 28th September, 1950

Application of Information Theory to a Study of the Sense Organs and the Central Nervous System.

Chairman - Professor le Gros Clarke, F.R.S.

9)	Communication Theory and Linguistic Theory	Dr. D. B. Fry, Phonetics Dept., University College, London.
10)	Hearing	T. Gold, Cavendish Laboratory, Cambridge.
11)	The Problem of the Information which the Brain Receives from the Eye	Dr. W. A. H. Rushton, F.R.S., Physiological Laboratory, Cambridge.
12)	Information Theory in Psychology	Dr. W.E. Hick, Psychology Laboratory, Cambridge.

/Chairman

Chairman - Professor E. G. Adrian, O.M., F.R.S.

13)	Possible Features of Brain Function and their Imitation	Dr. W. Grey Walter, Burden Neurological Institute, Bristol.
14)	Significance of Information Theory to Neurophysiology	Dr. J. A. V. Bates, National Hospital for Nervous Diseases, London.
15)	Information, Machines and Brains	Dr. A. M. Uttley, T.R.E., Malvern.
16)	Statistics for the Chess Computer and the Factor of Mobility	Dr. Eliot Slater, National Hospital for Nervous Diseases, London.

Friday, 29th September, 1950

Chairman - Professor J. L. van Soest.

17)	Criteria of Prediction and Discrimination	Dr. J. H. Westcott, Electrical Engineering Dept., Imperial College, London.
18)	Entropy, Time and Information	D. M. MacKay, King's College, London.

Chairman - Professor Willis Jackson.

19) Concluding discussion, opened by a talk by Professor J. B. Wiesner on the work of the Electronics Laboratory of the Massachusetts Institute of Technology.

PARTICIPANTS

VISITORS FROM ABROAD

FRANCE	A. Fromageot, Laboratoire Central de Telecommunications, Paris.
	M. D. Indjoudjian, Ing. des Postes, Telegraphes et Telephones, Paris.
	J. R. V. Oswald, Ing. de Telecommunications, Paris.
	M. Steinberg, Ecole Normale Superieure, Paris.
	J. Ville, Ing. Societe Alsacienne de Constructions Mecaniques, Paris.
	Prof. P. Grivet, Laboratoire de Radio-électricité, Paris.
GERMANY	Dr. F. A. Fischer, Central Laboratory - German Post Office, Darmstadt.
	Dr. H. Holzwarth, Dr. Ing. Systems Development, Central Research Laboratories, München.
	Dr. W. Meyer-Eppler, Phonetics Institute, University of Bonn.
HOLLAND	Professor G.H. Bast, Central Laboratory, P.T.T., The Hague.
	N. Rodenburg, Philips Telecommunications, Hilversum.
	Dr. J. F. Schouten, Philips Telecommunications, Hilversum.
	Professor J. L. van Soest, Technical University, Delft.
	Dr. F.L. Stumpers, Philips Research Laboratories, Eindhoven.
NORWAY	Professor H. Dahl, Chr. Michelsens Institute, Bergen.
	N. Knudtzon, Norwegian Defence Research Establishment, Bergen.
SWEDEN	C. G. Aurell, Messrs. Ericsson, Stockholm.
	S. Comet, Swedish Defence Staff, Bromma.
	Professor B. Haard, Royal Institute of Technology, Stockholm.
	N. H. Lundquist, Defence Research Institute, Stockholm.
SWITZERLAND	Prof. B. van der Pol, C.C.I.R., Geneva.
	Dr. Y.Y. Mao, C.C.I.R., Geneva.
	K. Willi, Swiss General Post Office, Brunnenweg.
UNITED STATES OF AMERICA	Dr. Bell, Scientific Liason Office, U.S. Embassy, London.

BRITISH UNIVERSITY and ALLIED DEPARTMENTS

H.B. Barlow, Physiological Laboratory, Cambridge.
G. A. Barnard, Mathematics Department, Imperial College, London.
R. L. Beurle, St. Dunstans, London.
A. H. Boothroyd, Electrical Engineering Department, Imperial College, London.
K. G. Budden, Cavendish Laboratory, Cambridge.
G. D. Dawson, Neurological Research Unit, London.
P. Denes, Phonetics Department, University College, London.
F. Foster, Mathematics Department, University of Oxford.
I. J. Good, London.
Professor A. V. Hill, Biophysics Department, University College, London.

/S. J. Holt,

S. J. Holt, Nature Conservancy, Edinburgh.
Professor C. Holt Smith, Military College of Science, Shrivenham.
Z. Jelonek, Polish University College, London.
D. G. Kendall, Magdalen College, Oxford.
Professor M. G. Kendall, London School of Economics.
R. Kompfner, Clarendon Laboratory, Oxford.
Professor A. Lee, Military College of Science, Schrivenham.
P. A. Merton, Neurological Research Unit, London.
P. A. Moran, Institute of Statistics, Oxford.
Professor M. H. A. Newman, Mathematics Department, Manchester University.
R. C. Oldfield, Institute of Experimental Psychology, Oxford.
J. W. S. Pringle, Neurophysiology Department, Cambridge.
F. Roberts, Anatomy Department, University College, London.
W. Ross Ashby, Barnwood House, Gloucester.
H. W. Shipton, Burden Neurological Institute, Bristol.
D. A. Sholl, Anatomy Department, University College, London.
Professor J. Thomson, Royal Naval College, Greenwich.
K. D. Tocher, Mathematics Department, Imperial College, London.
A. M. Turing, Mathematics Department, Manchester University.

BRITISH GOVERNMENT ESTABLISHMENTS

Admiralty.	W. P. Anderson. H. C. Calpine. J. Swaffield. S. Vajda.
General Post Office.	Dr. W. G. Radley. R. J. Halsey. W. J. Bray. N. W. Lewis. W. E. Thomson. G. Timms.
National Physical Laboratory, D.S.I.R.	Dr. E. C. Bullard. F. M. Colebrook. J.C. Evans. E. C. Fieller. J. E. L. Michel.
Radar Research and Development Establishment, Ministry of Supply.	A. E. Bailey. J. Brown. P. H. Blundell.
Radar Training Battalion, R.E.M.E.	E. Jeffery.
Radio Research Station, D.S.I.R.	R. E. Burgess. W. R. Piggott.
Royal Aircraft Establishment, Ministry of Supply.	G. F. Clarke. R. A. Fairthorne. D. G. Reid.
Signals Research and Development Establishment, Ministry of Supply	R. H. Barker. E. Fitch. E. V. Glazier. D. Williams.

/Telecommunications

Telecommunications Research Establishment, Ministry of Supply	J. F. Atherton. I. L. Davies. S. Jones. B. Newsam.
British Broadcasting Corporation	H. L. Kirke. W. Proctor Wilson. P. A. T. Bevan. A. M. Beresford-Cooke. S. H. Padel.

BRITISH INDUSTRIAL CONCERNS

Albright & Wilson Ltd.	F. W. Masham.
Boulton Paul Aircraft Ltd.	A. Bruce.
British Telecommunications Research Ltd.	J. Lawton.
British Thomson-Houston Co. Ltd.	D. J. Mynall.
Electric and Musical Industries Ltd.	W. J. Percival. T. J. Rey.
Ferranti Ltd.	J. B. Smith. D. F. Walker.
General Electric Co. Ltd.	J. E. Bryden. M. Levy. S. H. Moss. L. Stenning.
Marconi Co. Ltd.	C. D. Colechester. E. Eastwood. S. Millington.
Standard Telephones and Cables Ltd.	K. G. Hodgson. L. C. Pocock. A. H. Reeves.

NATURE, SCOPE AND TERMINOLOGY OF INFORMATION THEORY

COMMUNICATION THEORY, PAST, PRESENT AND PROSPECTIVE

by
D. Gabor

Modern communication theory originated from the attempts of communication engineers to understand what they were doing, in the most general terms. When the mathematical concept of information had crystallized out of these attempts, it was found that a track beaten by theoretical physicists led to the same idea, and, as usual, it was found that pure mathematicians had prepared the way to even more general developments. The new science of information theory thus connects several fields of research, old and new, each with its own techniques.

The scope of communication engineering.

Communication is the transmission of information from mind to mind. The communication engineer is directly concerned only with a part of this process, say from microphone to earpiece, or from television camera to the screen of the cathode ray tube. As an introduction to the wider applications of information theory let us first consider communication in this restricted sense.

Completeness and economy.

When electrical communications started the first requirement was completeness. Morse had to transmit the whole alphabet; Graham Bell the whole range of speech sounds. But the question of economy also came in, almost from the start. It was soon found that it was necessary, but also sufficient, to make the transmission line, and the terminal equipment, responsive only to those frequencies which the human voice apparatus can produce, and which can be heard by the human ear. Curiously it took much longer before the importance of bandwidth was realized also in telegraphy. But on the other hand it was in telegraphy where the advantages of an economic coding were first recognized. Though the inventors were led only by their sound common sense, the Morse code is almost ideal for a system which associates letters in the English language with dots and dashes, and the Baudot code is in fact ideal for messages of which we know nothing in advance except that they use the 32 letter alphabet. The modern theory of coding, due chiefly to Shannon, can offer better advice if the statistical structure of the message is known, e.g. if all messages are sent in plain English. In this case one can devise codes, operating not with single letters but with digrams, trigrams etc. which are about twice as efficient. It is somewhat doubtful, however, whether these will be much used in practice. The redundancy of the English language is a great help in reconstructing imperfectly transmitted messages. In a perfect code a mistake could not be bridged by guessing. It is possible that a compromise can be arrived at, but it is evident that telegraphy can obtain only a moderate benefit from the application of modern communication theory.

Coding

Ideal coding.

Limited application in telegraphy.

Great potential use in telephony.

In telephony the situation is very different. In telegraphy a letter is transmitted in the Baudot system by five "binary selections" or five "bits". (Short for "binary digits".) In a telephone channel on the other hand, with a bandwidth of 3000 cycles and a signal : noise ratio of 30 decibels the number of bits/second is $2 \times 3000 \times 3.32 \times 3 = 60{,}000$, yet even in rapid speech the number of speech sounds per second is only about 15: thus it takes about 4000 bits to transmit a letter instead of 5. Of course the telephone transmits more than mere letters; it conveys the individuality of the speaker and the emotional content of his message, but even allowing for this there remains a large untapped source of potential gain. But if we want to realize this, we must go a little beyond the narrow definition of communication engineering. It is necessary to include the mind of the speaker and his voice organs, or at least the ear plus brain of the

receiver in the communication chain. It is not necessary, however, at least in principle, to go into the physiology or into the psychology of speech, hearing and understanding. The human receiver can be considered, as is done in experimental psychology as a "transducer" which responds differently to certain stimuli, without knowing anything about its internal operation. The problem is to map those physical characteristics of speech signals which give distinguishable or significantly different responses. This is a vast programme, in which only modest beginnings have been made so far.

Speech recognition.

No band-saving to be expected in music transmission.

While telephony is a hopeful field, the results which may be expected from the application of communication theory to broadcasting must remain, in the best case, about as modest as in telegraphy. The human ear is an almost half-ideal instrument, and polyphonic music utilises its capabilities to the limit. But even the modest bandwidth-saving of about one half will probably never be realised, because of the prohibitive price of the receiving equipment, which would have to be very complicated.

Transmission systems as coding methods.

There is, however, the rather different question of improving the quality and freedom from noise and interference in broadcasting by a judicious wasting of waveband, as practised in Frequency Modulation. Here modern communication theory has given valuable quantitative criteria to decide what appeared formerly as a battle of opinions. Taking the word in a more general sense than before, all the various transmission systems such as AM, FM, PAM, PCM, etc. are samples of different coding. Once again the sound sense of inventors was vindicated when it was shown that Pulse Code Modulation, (PCM), with a sufficiently complicated code, strikes an almost ideal bargain in trading frequency band against signal : noise ratio.

PCM

Applications of communication theory to problems of physiology and psychology.

We have seen already, that even practical communication engineering cannot leave the human "terminal" out of its considerations. It is an evident temptation to step over the limit, but we must ask first what communication theory can offer when applied to the old problems of the sensory organs, the nervous system and the mind? The answer is twofold. In the first line, communication theory brings to bear on the problem certain general principles, which must apply to the unknown structure of the nervous system as well as to man-made communication systems. But this contribution must not be overrated. The general principles in question are either mathematical theorems, or general physical principles, and as such by no means unknown to the numerous trained physicists and mathematicians who have devoted themselves in the past to problems of physiology and of experimental psychology. Far more important is the second kind of contribution; the wealth of techniques for solving communication problems, which might fertilize physiological research by analogy. The crucial step in almost every scientific discovery is the substitution of a model for the thing investigated, which has a similar functional behaviour. The process of guessing is greatly facilitated if there are already man-made devices for a similar purpose, but if not, they can be invented in a field in which human thought has already learned to move easily. As an example, looking for a safe method of telemetering, an engineer of the Western Electric Co. invented Pulse Rate Modulation at just about the same time as Adrian proved that nerve impulses are in fact transmitted by PRM. In this case the electrical invention came a little late to help the physiologist, but it may be different in the future. Much effort of the best brains has gone lately into devising complicated calculating machines, which perform some of the "higher" functions of the human intellect. It may well be that some of the real structural features of the brain are already embodied in these ingenious computers.

Techniques more important than general principles.

Calculating machines and the brain.

New inventions
The automatic typist
Translating machines
General robots.

Sensory replacement.

It is likely that the invention of complicated computers and of other machines for performing difficult "intellectual" operations will itself take a great step forward under the impact of communication theory and communication techniques. Chess playing and sonnet composing machines may not be of great practical importance, and it may be fervently hoped that George Orwell's "versificator" will never be built, but the "automatic typist" which takes down dictation, and the "translating machine" which saves the trouble of learning foreign languages may be more desirable, at least from a purely commercial point of view. We must also mention the "universal robots" forecast not only by Karol Capek but also by N. Wiener in his "Cybernetics", machines which will react with as much intelligence as any routine worker to a foreseen set of stimuli. A far less sinister field of application is the replacement of sensory organs, which has made already very satisfactory and remarkable progress.

Communication theory and the foundations of science: Information theory.

So far we have dealt with communication theory, in which information is supposed to be pre-existent in some mind, or at least foreseen to such an extent that a pigeonhole or a response are provided for it when it emerges. But the concept of Information has wider technical applications than in the field of communication engineering. Science in general is a system of collecting and connecting information about nature, a part of which is not even statistically predictable. Communication theory, though largely independent in origin, thus fits logically into a larger physico-philosophical framework, which has been given the general title of "Information Theory". It has already made some progress, and has made contact with formal logic and the mathematical theory of representation on the one hand, with epistemology on the other. Again some caution is needed when appraising the prospects of this young subject. It may be remembered, that in the past physics had less to learn from philosophy than philosophy from physics, though it is usually the philosopher who likes to have the last word. But information theory is of course not all philosophy, and some of its quantitative concepts may well prove fruitful in the older sciences, not only in the new ones.*

The present Symposium is itself an experiment in the transmission of information, under rather unprecedented circumstances. No statistical prediction of its effects is possible, except that it may be hoped that it will increase the sum total of information among its participants rather than decrease it.

---

* A glossary relevant to this most general aspect of Information Theory by Mr. D.M. Mackay appears on pages 10-21.

GLOSSARY OF PHYSIOLOGICAL TERMS

by

J.A.V. Bates

1. **The Nervous System** is a collection of cells specially adapted for the propagation of electrical impulses. Such cells are distinguishable in the sea-anemonae, and in all higher forms of life. By means of this conducting system a variety of disturbances in the external or internal environment of the animal can affect its motor mechanisms in a way which tends to maximise the chances of survival of itself and its species. In general motor activity tends to nullify the effects of sudden changes in environment. This attribute of living matter has been termed the principle of .....

2. **Homeostasis**

In the higher animals (vertebrates) it is convenient to distinguish two divisions of the nervous system; first .....

3. **The Central Nervous System,** which comprises all those nervous elements (defined below) contained within the bony frame-work of the skull and spinal cord; and secondly .....

4. **The Peripheral Nervous System** which comprises all those nervous elements outside the skull and spinal cord.

The basic unit of the nervous system is .....

5. **The Nerve Cell,**
6, 7 **(Neuron, Nervous Element)** which was first clearly defined in about 1890. It is found in a large variety of shapes and sizes, but in general the following parts can be distinguished: .....

8. **The Cell Body** which contains the cell nucleus; .....
9. **Dendrites,** which are multiple short processes arising from the cell body, and the .....
10. **Nerve Fibre**
11, 12 **(Axis Cylinder, Axon)** which is one process from the cell body appreciably longer than the dendrites.

The cell body will have a diameter between 10μ and 200μ (μ = 1/1000 mm.). The dendrites may be 5μ to 1 cm. in length, and 3/4μ to 20μ diameter. The nerve fibre may be 20μ to 2 metres in length, and 1μ to 20μ diameter.

13. **A Nerve Impulse**
14. **(Action Potential)** is identifiable as a disturbance of an ionic equilibrium which normally exists across a thin membrane surrounding the nerve fibre. The impulse has a .....
15. **Conduction Velocity** of from 1 to 10 metres/sec. This velocity depends to a large extent on the minute construction of the fibre, and to a lesser extent on fibre diameter, external temperature and other factors. Recovery of the surface membrane potential can be sufficiently rapid to permit a .....
16. **Frequency of Conduction** of up to 400 impulses/sec. The impulses in any nerve fibre can travel either away from or towards the cell body, but in life so far as we know, the impulse always travels in the same direction in any one fibre.

The impulse passes from one fibre to the next via a .....

The impulse passes from one fibre to the next via a .....

. Synapse.
. (Interneurone)

This may or may not exist whenever two or more fibres are separated by less than about 100μ. There is no microscopic appearance which certainly defines its presence, its existence can be inferred by functional tests which will be referred to later.

The nervous system is to some extent divisible into a .....

. Sensory System,
. (Afferent System)

and a .....

. Motor System
. (Efferent System)

but these two categories usefully include only a small proportion of the total. Separate nerve cells are used to provide a sensory inflow, and a motor outflow.

The sensory inflow comes from .....

. Sensory Cells
. (Receptors)

which can originate an impulse in a nerve fibre adjacent to them. The necessary cause for this impulse is a sufficiently sudden change in the cell's physical environment.

An aggregate of similar sensory cells comprise a .....

. Sense Organ

All varieties of sense cell are physically sensitive in one of three ways; a) Radiation (heat), b) Chemical Structure (taste) and c) Physical Impact (touch). The process of evolution has highly developed these sensitivities, so that the complete animal can now percieve intense heat at a distance (light), minute concentrations of particular molecules (smell) and touch at a distance (hearing). An impulse can be originated in nerve fibres (as distinct from sense cells) by artificially disturbing their surface membrane potential. This can be done for example by an electric current or injury, and it provides a technique for examining the nervous system.

The motor outflow has two primary divisions;

. The Somatic Motor System

which causes contractions of all muscles attached to the bony skeleton, and .....

. The Autonomic Nervous System

which activates contractile tissue in various organs of the body (gut, blood vessels etc.) The autonomic system is further subdivided into the .....

. Sympathetic System,
. Parasympathetic System.

and the .....

Aside from these systems particular aggregates or nerve cells and parts of nerve cells can be identified by naked eye appearance, and selective chemical stains (histology). Thus the brain has an outer covering largely of .....

. Grey Matter

containing cell bodies, and dendrites; and beneath it a layer of .....

. White Matter

which contains only nerve fibres.

. Tracts
, 34. (Bundles, Pathways)

are identifiable collections of nerve fibres within the Central Nervous System running together between the same destinations. About one hundred tracts have been identified and named within the human nervous system.

The brain is divisible into a hind-brain, mid-brain and fore-brain. Study of its evolutionary history strongly supports the idea that the .....

35. Fore-Brain
36. (Cerebrum)

is the most recent acquisition, and is primarily concerned with operations on data from the animal's specially sensitive "distance" receptors (24). The primate fore-brain is distinguished from others by the large amount of it devoted to vision. The fore-brain comprises mainly the .....

37. Cerebral Hemispheres

whose folded light grey outer covering is termed the .....

38. Cerebral Cortex.

The cortex is further divided into the visual c., sensory c. and motor c. etc. on functional and anatomical grounds.

The microscopic appearance of thin stained sections of brain is that of inextricable three dimensional tangle of nerve cells and fibres in a variety of shapes and sizes. A gross regularity of pattern can in some places be discerned under low powers of magnification. One school believes that this complexity can be broken down into a repetition of only two basic patterns. A

39. Multiple Chain Circuit,

whereby a single impulse in one fibre can lead to a single impulse in a number of other fibres. and a .....

40. Closed Chain Circuit

whereby a single impulse in one fibre can lead to a succession of impulses in another single fibre. It is necessary to state however that the validity of this simplification is not at the moment generally accepted.

Various terms are widely current as a result of experimental studies of the behaviour of the whole organism (psychology), and of the functioning of individual systems comprising the organism (physiology). It is crucial to such studies that the separation of the nervous system into a sensory and a motor side implies that the impulse in a sensory fibre can never give rise to an impulse in a motor fibre without passing across at least one synapse - more usually two or more synapses are involved, in which case conduction of the impulse is continued along one or more

41. Internuncial Fibres.

Conduction across one synapse involves a .....

42. Synaptic Delay

which is $\frac{1}{2}$ to 1 m.sec. This fact, combined with accurate measures of conduction rate, enables the number of synapses in a sensory-motor pathway, also called a .....

43. Reflex Arc,

to be estimated. The particular synapse involved in a reflex-arc cannot however be discovered, only its general location. The important consequence of this arrangement of synapses and internuncial fibres is that a motor output is never completely specified by a particular sensory input, but is the result of the .....

44. Integration

of particular sensory data with other and more general sensory data. In the limit this additional data may result in .....

45. Inhibition

of the usual response to the stimulus. As Sherrington has said, "at the synapse, function fluctuates".

The animal is born with a number of reflex arcs, i.e. with a number of pre-formed stimulus-response elements, whose operation together produce a characteristic pattern of behaviour. They vary from the simple to the exceedingly complex (e.g. motor activities of the new-born). An example of a simple inborn reflex system is that which provides .....

46. Reciprocal Innervation, which refers to the relaxation of a muscle when another muscle which is in mechanical opposition (antagonistic) to it is actively contracting. The probability of a particular pattern of motor activity following a particular stimulus is constantly being modified during the life of the animal as a result of its "memory" of the past. When a consistent modification of behaviour has come about as a result of a specific stimulus-complex the animal is said to possess a .....

47. Conditioned Reflex to that stimulus.

If electrodes are placed on the skull and connected to a suitable high gain amplifier there is evidence of constant electrical activity from the brain beneath. The potential changes are small - rarely more than 50μ volts - partly owing to the amount of conducting tissue shunting the source. From the human brain there is, in a large proportion of the population, but not invariably, a characteristic oscillation of potential at a frequency of about 10/sec. This is usually most marked over a region of the cortex known to be associated with vision, though it may be widespread. It is known as .....

48. The Alpha Rhythm.
49. (Berger Rhythm)

It disappears and is replaced by potential changes without characteristic periodicity when the attention of the subject is directed visually. For example it appears best when the eyes are shut, but it can be made to disappear under this condition by the effort of visualising a remembered scene.

---

24. Action Potential
20. Afferent System
48. Alpha Rhythm
27. Autonomic Nervous System
11. Axis Cylinder
12. Axon
49. Berger Rhythm
33. Bundle
8. Cell Body
3. Central Nervous System
37. Cerebral Hemispheres
38. Cerebral Cortex
36. Cerebrum
40. Closed Chain Circuit
47. Conditioned Reflex
15. Conduction Velocity
38. Cortex
9. Dendrites
22. Efferent System
35. Fore-brain
16. Frequency of Conduction
30. Grey Matter
2. Homeostasis
45. Inhibition
44. Integration
18. Interneurone
41. Internuncial fibre
21. Motor System
39. Multiple Chain Circuit
5. Nerve Cell
10. Nerve Fibre
13. Nerve Impulse
7. Nervous Element
1. Nervous System
6. Neuron
29. Parasympathetic System
34. Pathway
4. Peripheral Nervous System
24. Receptor
46. Reciprocal Innervation
43. Reflex arc
28. Sense Organ
23. Sensory Cell
19. Sensory System
26. Somatic Motor System
28. Sympathetic System
42. Synaptic Delay
17. Synapse
32. Tracts
31. White Matter

THE NOMENCLATURE OF INFORMATION THEORY

by
D.M. MacKay

(A) FOREWORD

This is not a glossary in the sense of an agreed list of standard terms in Information Theory. No such agreement yet exists in this new subject. It is rather an attempt to collect and collate as many as possible of the terms which are in current use, and to define tentatively the ways in which they are related and the senses in which they may be interpreted without conflict. Only time will show whether these interpretations will be adequate or acceptable; but if it succeeds only in demonstrating the complementary and non-competitive relationship of different current approaches to the problem this somewhat Augean task will have been worth attempting.

The glossary is not of course meant as an "introduction for beginners", so much as a logical framework in which terms may be seen in perspective as they arise. The proportions of space devoted to different aspects of the subject are therefore no indication of their relative importance, being dictated merely by the exigencies of exposition. Although the approach is bound to be a personal one, a considerable effort has been made to base it on logical consideration of what has actually been meant by various authors, in consultation as far as possible with other speakers. The particular terms "metrical", "structural" and "selective" information may possible be thought unsuitable; but the logical distinctions for which they stand appear to be essential. Confounding of these is believed to be the source of much unedifying debate.

The reader will realise that the width of subject-matter embraced in the opening paragraphs of section (C) is representative only of the logical purview of the subject as a whole, and is intended neither as a forecast of the scope of the present Symposium, nor as a territorial raid on the established fields through which the concepts of Information Theory define a unifying path.

(B) INTRODUCTORY: WHAT INFORMATION THEORY IS ABOUT

In everyday speech we say we have received Information, when we know something that we did not know before: when 'what we know' has changed. If then we were able to measure 'what we know', we could talk meaningfully about the amount of information we have received, in terms of the measurable change it has caused. This would be invaluable in assessing and comparing the efficiency of methods of gaining or communicating information.

Information Theory is concerned with this problem of measuring changes in knowledge. Its key is the fact that we can represent what we know be means of pictures, sentences, models or the like. When we receive information, it causes a change in the symbolic picture or representation which we would use to depict what we know. It is found that changes in representations can be measured; so 'amount of information', actually in more than one sense, can be given numerical meaning. It is as if we had discovered how to talk quantitatively about size, through discovering its effects on measuring-apparatus. We should at once find that it had the quite different but complementary senses of volume, area, and length - if not others. The analogy is potentially misleading, but may show us what to expect.

Right at the start the term Information takes on two different kinds of meaning in answer really to different kinds of question. An example will illustrate the point. Two people A and B are listening for a signal which they know will be either a dot or a dash. A dash arrives. A makes various measurements, represents what has happened by a graph, and asserts that there was "a good deal of information" in the signal. B says "I knew it would be either a dot or a dash. All I had to do was to choose one or other of those prefabricated representations. I gained little information."

A and B are not of course in disagrement. For lack of a vocabulary, they are using the same word 'information' as a measure of different things. A is using the term in the sense of what we call 'Scientific Information'. This in itself has two aspects, relating roughly to the number of independently-variable features (structural information) and the precision or reliability (metrical information) of the representation he has made. The knowledge which he says has increased, is knowledge of what has actually happened and been observed.

What of B? He was not waiting to observe everything that happened. He already knew that for his purposes only two kinds of representation would be needed, and he had prefabricated one of each. The knowledge he acquired was knowledge of which representation to select. B was therefore using 'information' in the sense of 'that which determines choice', which we may call selective information.

One word which was unexpected would yield B more selective information than a whole message which he was already sure he would receive.

A's approach is typical of the physicist, who wants to make a representation of physical events which he must not prejudge. B's is typical of the communication engineer, whose task is to make a representation at the end of a communication channel, of something he already knows to be one member of a set of standard representations which he possesses. His concern is therefore not with the size or form of a representation, but with its relative rarity, since this will govern the complexity of the "filing-system" he should use to identify it. Each however may on occasion find both approaches relevant to different aspects of his work.

To sum up, if we ask how much information there is in a given representation, we may mean: "How many distinct features has it? How many elementary events does it describe?", in which case we require answers in terms of amounts of Scientific information; or we may be ignoring questions of the size and complexity of the representation, and thinking instead of the complexity of the selection-process by which it was identified, meaning: "How unexpected was it? How small a proportion of all representations is of this form? In how many steps were you able to identify it in your 'filing-cabinet' of possibilities?". In this case our question refers to amount of selective information. Rarity here is the touch-stone, as against logical structure in the first case. It will be realised - and this may be an important help to the understanding of the subject - that the term 'information' means something quite distinct from 'meaning'. If the reader begins by divorcing the two completely, he may find it easier to trace the connections in any subsequent reunion.

Our purpose in the explanatory glossary which follows is to show how the terms which have arisen in these different connections are related, and to scotch if possible once and for all any suggestion that these complementary senses of the term Information are in any way competitive.

(C) EXPLANATORY GLOSSARY

(1) The Scope of Information Theory

(1.1) Information Theory is concerned with the making of representations - i.e. symbolism in its most general sense.

Representations

(1.1.1) By a representation is meant any structure (pattern, picture, model) whether abstract or concrete, of which the features purport to symbolize or correspond in some sense with those of some other structure.

(1.1.2) The physical processes concerned in the formation or transformation of a representation are thus distinguished from other physical processes by the element of significance which they possess when conceived as representing something else.

(1.1.3) For any given structure there may be several equivalent representations, defined as such through possessing certain abstract features in common.

(1.2) It is these abstract features of representations which are of interest in Information Theory. Its aims are (a) to isolate from their particular contexts those abstract features of representations which can remain invariant under reformulation,* (b) to treat quantitatively the abstract features of processes by which representations are made, and (c) to give quantitative meanings to the several senses in which the notion of amount of information can be used.

(1.3) The scope of Information Theory thus includes in principle at least three classes of activity:

(1.3.1) Making a representation of some physical aspect of experience. This is the problem treated in Scientific Information-Theory.

Scientific Information-Theory

(1.3.2) Making a representation of some non-physical (mental or ideational) aspect of experience. This is at the moment outside our concern, being the problem of the Arts.

(1.3.3) Making a representation in one space B, of a representation already present in another space A. This is the problem of Communication Theory, B being termed the receiving end and A the transmitting end of a Communication Channel.

Communication Theory, Communication Channel.

(1.3.3.1) By a space is meant any physical or abstract mathematical coordinate - framework or manifold, of any number of dimensions, in which the elements of a representation can be ordered.

Space

(1.4) These categories are not of course exclusive. The problem of communication in particular is seldom separable from one or both of the first two. In its present state of development Information Theory is concerned mainly with (1.3.1) the problem of representing the physical world, and (1.3.3) the problem of communicating representations (of any kind). It is communication theory (1.3.3) however, with its immense practical importance, which has received the greatest attention; and it is only the logical priority of the other two which prevents it from coming first on the list.

(1.5) The situation confronted by Information Theory is summarised in the somewhat naive diagram of Fig.1, on the basis of the conventional view that the task of Physics is to describe a world of physical events most of which must be inferred from instrumental observations. Instrumental displays are therefore not merely physically observable structures, but representations. But they are not models; as is shown in Fig.1, a model can only appear after an abstract conceptual stage. The details of the communication-process A - B are not shown. Normally B also has access to the physical world but the sequence chosen is merely illustrative of essential order.

---

*This aspect of the subject is already an established branch of mathematics under the name of Representation Theory or Abstract Group Theory.

The "World of ideas" of course occupies a purely symbolic position in the pattern and is not thereby localised or objectivised.

(2) Information

The foregoing definition of the scope of Information theory provides the necessary background for the definition of Information.

Information

(2.1) In all its senses, the term can be covered by a general 'operational' definition (i.e. a definition in terms of what it does: as (e.g.) force is classically defined in terms of the acceleration which it causes or could cause.) The effect of information is a change in a representational construct.

(2.1.1) Information may be defined in the most general sense as that which adds to a representation.

(2.2) This leaves open the possibility that information may be true or false.

(2.2.1) When a representation alters, we define the new information as true if the change increases the extent of correspondence between the representation and the original.

(2.2.2) The information is said to be false if the change diminishes the extent of this correspondence.

(3) Measurement of Information

Two quite different but complementary approaches are possible to the measurement of Information, and have given rise to quite different senses of the term:

(3.1) From a quantitative analysis of what a representation portrays, we can isolate fundamental numerical features common to all its equivalent representations, and can say that they constitute the 'corpus of information' which it contains or represents.

(3.1.1) "Amount of information" in this context is a measure of complexity.

(3.2) But if instead of asking "How many elements etc. are there in this representation?", we ask "In how many stages, and in what way, has it been built up?", we may arrive at a different kind of measure. If we consider that the same representation could be built up in a number of different ways according to the amount of prefabrication used, it is evident that this measures something different from the complexity of the pattern.

(3.3) In the following paragraphs four and five, these two approaches will be discussed. The first has given rise to two complementary definitions of what has been called "amount of information", and the second to a third. These, however, are not rivals, but are autonomously valid measures, appropriate in answer to different questions.

(4) Analysis of representations

(4.0.1) Representations communicable in a two-valued (yes-or-no) form are necessarily quantal in structure, since an "imperceptible change" in a two-valued logical form is by definition meaningless. All the changes are discrete, therefore the elementary concepts of logical representations are discrete and enumerable.

(4.0.2) In fact a large class of such representations can be reduced to a form made up only from identical elements, so simple that their only attribute is existence. This fact provides a basis for the quantitative analysis of such representations. (Representations not amenable to precise logical description have not so far been considered in the theory, though a large class of these might be handled in terms of approximate quantal equivalents representing "upper and lower bounds" to their logical content.)

(4.0.3) In general a pattern reduced to such fundamental terms will contain a certain number of distinguishable groups or clusters of elements, the elements in each group being indistinguishable among themselves. There are thus two numerical features of interest.

(1) The number of distinguishable groups or categories in a representation, and

(2) The number of elements in a given group or category.

(4.0.3.1) The number of groups, and the numbers of their elements, may be thought of as respectively analogous to the number of columns and the number of entries per column in a histogram.

(4.1) Structural Information

Structural Information-Content

(4.1.1) The number of distinguishable groups or categories in a representation - its dimensionality or number of degrees of freedom or basal multiplicity - is termed its Structural Information-Content.

Logon

(4.1.2) The unit of structural information, one logon, is that which enables one new distinguishable group or category to be added to a representation. Thus structural information is not concerned with the number of elements in a pattern, but with the possibility of differentiating between them.

(4.1.2.1) For example if we are counting identical sheep jumping a gate and have no sense of time, our result can only be represented by a certain total number; but if we have a clock, we can now define what we mean by "the number in the first minute" and "in the second" and so forth, and represent our result by a set of distinguishable sub-totals. The clock has provided structural information. In a similar way the ability to distinguish (e.g.) spatial position would provide distinguishable sub-totals and hence structural information.

Logon-content

(4.1.3) Logon-content is a convenient term for the structural information-content or number of logons (number of independently variable features) in a representation (e.g. the number of coefficients required to specify a given waveform).

Logon-capacity (Possible synonym, logon-density)

(4.1.4) The number of logons provided by apparatus per unit or coordinate - space (centimetre, square centimetre, second, etc.) is terms its logon-capacity.

(4.1.4.1) For example, a channel whose bandwidth permits of ℓ independent readings per second has a logon-capacity of ℓ ; in a microscope, logon-capacity is a measure of resolving-power; and so forth.

Structural scale-unit

(4.15) The reciprocal of logon-capacity is termed the structural scale-unit for the apparatus.

(4.2) Metrical Information

Metrical information content

(4.2.1) The number of logical elements in a group or pattern is termed its Metrical Information-Content.

Metron

(4.2.2) The unit of metrical information, one metron, is defined as that which supplies one element for a pattern. Thus the amount of metrical information in a pattern measures the weight of evidence to which it is equivalent. Metrical information gives a pattern its weight or density - the 'stuff' out of which the 'structure' is formed.

(4.2.3) Each metrical unit may be thought of as associated with one elementary event of the sequence of physical events which the pattern represents.

Metron-content (syn: Metrical Information Content)

(4.2.4) Thus the amount of metrical information in a single logon, or its metron-content can be thought of as the number of elementary events which have been subsumed under one head or 'condensed' to form it. For example in the case of a numerical parameter, this is a measure of the precision with which it has been determined.

(4.2.4.1) Notice that these elements are indistinguishable, so that their number is not the number of binary digits (5.1.3) to which the logon is equivalent.

(4.2.5) When we come to represent the results of physical observations, we are often interested in magnitudes which are not directly proportional to the metron-content. The representations we use do not then show the metron-content explicitly. It must be clearly realised that metron-content as defined is a measure of the number of elements appearing when what is believed to have happened is represented in its most fundamental physical terms.

Conceptual Scale

(4.2.6) Thus if an estimate is made of a parameter from a statistical sample, the elementary events concerned are the arrivals of 'unit-contributions' to the sample. These could be represented by the number of intervals occupied on a conceptual scale proportional to metron-content.

Proper-scale

(4.2.7) On the other hand the usual representation shows the magnitude of the parameter concerned, and not generally the metron-content, on a linear scale, graduated in elementary intervals which in the useful limit are just large enough to give the representation of the magnitude (scale-reading) a probability of $\frac{1}{2}$; Such a scale is termed a proper-scale. Now the probable error, in a normal population, is inversely proportional to the square root of the size of the sample. Hence the magnitude in such a case would be shown as occupying on the proper scale a number of elementary intervals proportional only to the square root of the number of elementary events. The number of metrons is (in this special but common case) the square of the number of occupied intervals shown in this less fundamental physical representation.

Numerical energy

(4.2.7.1) In connection with radar information the term numerical energy has been used to represent what is essentially the metron-content of a signal. It is the ratio (Total Energy)/(Noise-power per unit bandwidth).

(4.2.8) In general then a clear distinction exists between (a) the fundamental representation on a conceptual scale showing the invariant number of logical elements, and (b) the representation of the magnitude which is of practical interest. In fact the connection between the two is little closer than that between the precision with which a given variable can be measured, and its magnitude. Precision increases monotonically with metron-content, but few quantities are linearly related to metron-content. Power and energy in the classical case are among the few exceptions; this accords with their apparently fundamental status among physical concepts.

Scale-unit

(4.2.9) The scale-unit of a magnitude is the minimum interval in terms of which the scale can usefully or definably be graduated.

(4.2.9.1) For a magnitude imprecisely known it is defined as above (4.2.7) to be equal to the probable range of error. A magnitude supported by a single metron occupies just one interval on such a scale. In practice larger units are often used - e.g. range of standard error.

Coincidence-relations

(4.2.9.2) But it should be remembered that in theoretical representations the size of the scale-unit is generally limited by our inability to define a smaller unit in terms of coincidence-relations out of which physical statements are constructed.

Metron-capacity
Syn: Metron-density

(4.2.10) The number of metrons per unit of coordinate-space is termed the metron-capacity or metron-density of a physical observation-system (Cf. 4.1.4).

Conceptual unit
Metrical scale-unit (undesirable Synonym)

(4.2.11) The coordinate interval in which one metron is acquired is termed a conceptual unit or (undesirably) a metrical scale-unit of coordinate.

(4.2.12) It will be noted that metron-content is necessarily positive.

(4.2.13) Returning to the example of the jumping sheep, let us now suppose that we are trying to determine a figure for the average value of some parameter of a sheep, in each group which we are able to distinguish. Assuming for the sake of illustration that the parameter is normally distributed, the metron-content of each estimate would be proportional to the number of sheep per group, and the probable error in each estimate would be inversely proportional to the square root of this number. Hence the number of proper-scale-intervals occupied by the estimated parameter would be proportional to the square root of the metron-content of each group.

(4.2.14) The term 'amount of information' in this metrical sense was first used by Fisher (Ref.Bib.) who defines it as follows: Suppose we have a probability distribution function $f(x,x_0)$ showing how in a given population a variable x is distributed about a parameter x_0, (e.g. $f(x,x_0) = Ae^{-\frac{(x-x_0)^2}{2\sigma^2}}$);

The amount of information in n samples from the population is defined as n times the weighted mean of $(\frac{\partial}{\partial} \frac{\log f}{x_o})^2$ over the range of x - i.e. n $\int_{-\infty}^{\infty} (\frac{\partial \log f}{\partial x_o})^2 f \, dx$.

Equivalent forms are

$$n \int_{-\infty}^{\infty} \frac{1}{f} (\frac{\partial f}{\partial x_o})^2 \, dx \text{ and } -n \int_{-\infty}^{\infty} (\frac{\partial^2 \log f}{\partial x_o^2}) f \, dx.$$

In the case of a normal distribution, this reduces to the reciprocal of the variance, provided that the range is independent of x_o. It is thus a direct measure of precision, though it is not dimensionless unless suitably normalised.

(4.3) Representation of Information

The Information content of a given representation is specified by setting down the metron-content of each logon. This may be represented in various ways.

Information-vector
Information-space

(4.3.1) One convenient method is to use a multi-dimensional information-vector in an information-space of which each axis represents one logon. The squares of the components of this vector are the metron-contents of the respective logons. Thus the square of the length of the vector itself is the total metron-content, the sum of the individual metron-contents.

(4.3.1.1) In this representation the angle between vectors has a direct interpretation as a measure of relevance. A dependent statement is defined by a ray in the space, and the metron-content afforded to it by the information is found by squaring the projection of the information-vector on the ray.

(4.3.1.2) A new complete representation may be set up by supplementing this dependent statement (4.3.1.1) by a set of others represented by orthogonal rays. This process amounts to a rotation of axes, which leaves us with a new total metron-content equal to the old.

(4.3.2.) The same processes can be represented in terms of matrix algebra, if the metron-contents of logons are set out initially as the elements of a diagonal Information-matrix. Dependent statements now define vector functions, and their metron-content is found by forming scalar products of the form $\phi'.I\phi$, where I is the information matrix, ϕ a vector function, and ϕ' its transpose. Under all complete transformations, the trace or sum of the diagonal elements of I remains invariant, being the total metron-content.

Trace (syn: Spur, Characteristic, Diagonal sum)

(4.3.3) An alternative geometrical representation suitable for some particular cases employs a three-dimensional histogram having a coordinate and its Fourier-transform (e.g. time and frequency) as its two basal axes. Since the number of logons provided by given apparatus is proportional to the product of bandwidth (q.v.) and conjugate coordinate, the base is divisible into equal cells each representing one logon. On each cell is erected a column having a height proportional to the logarithm of metron-content. This gives the total volume of the histogram the same qualitative significance as the logarithm of the volume spanned by the information-vector of 4.3.1.

(5) Communication: Replication of representations

(5.1) The problem of communication usually concerns representations of which all parts exist already in the past experience of the receiver. In other words the receiver already possesses prefabricated components of the representation.

(5.1.1) In fact it is generally assumed to be known that the complete representation to be replicated is one member of a finite ensemble (q.v.) of possible originals, some of which have in the past been received with greater frequency than others. We may then say that these more common messages "give less information" than the others, using the term 'information' in an important sense different from those so far mentioned.

Ensemble

(5.1.1.1) We are not here asking 'How big is it?' or 'How much detail has it?', but rather 'How unusual or unexpected is it?'. "How much trouble will it take to find it in my ensemble?".

(5.1.2) A convenient measure of information in this sense is the negative logarithm (base 2) of the prior probability of the representation concerned.

(5.1.2.1) The base 2 is chosen because a selection among a set of n possibilities can be carried out most economically by dividing the total successively into halves, quarters, eighths, etc. until the desired member is identified. The number of stages in this process is then the integer nearest to, and not less than, $\log_2 n$.

(5.1.2.2) The information measure so defined equal the number of independent choices between equiprobable alternatives which would have to be determined before the required representation could be identified in the ensemble of which it is a member. (The prior probability measures the fraction of the members of the ensemble which are of the required kind).

Selective information

(5.1.3) Information in the above sense of that which determines choice may be termed selective information.

Binary digit) syn: bit)

(5.1.4) The unit of selective information, one binary digit or bit, is that which determines a single choice between equiprobable alternatives.

(5.1.5) In a long sequence of different representations of which the i'th kind has a prior probability p_i, (and hence an average frequency of occurrence p_i) the average amount of selective information per representation is evidently the weighted mean of log p over all kinds of representation, or $H = -\sum p_i \log p_i$, which is also the standard definition of the entropy of a selection.

Entropy

(5.1.5.1) Where representations take the form of continuous functions, H takes the form $-\int f(x)\log f(x)\,dx$, where $f(x)$ is the probability distribution of the representative variable x. It is thus the weighted mean of $[-\log f]$ over the range of x.

(5.1.6) In practice the receipt of a communication signal disturbed by noise merely alters the form of $f(x)$, (generally narrowing it) and does not specify x uniquely. The amount of selective information received is then defined as the difference between the values of H computed before and after receipt of the signal.

(5.1.7) When two representative variables, x and y say, (discrete or continuous) are in question, knowledge of the value of one may affect the prior probability of the other. In the above notation, f(y) depends on x, so that the entropy H of y will also vary with x. If we picture an ensemble in which values of x occur in their expected proportions, we define the conditional entropy of y, $H_x(y)$ as the average value of the entropy of y (calculated for each value of x) over all members of this ensemble.

Conditional entropy

(5.1.7.1) This may also be described as the "weighted mean entropy" of y, weighted by the probability of getting the different values of x. It therefore measures our uncertainty about y when we know x.

(5.1.7.2) An analogous conditional entropy $H_y(x)$ can be defined for x when y is known.

(5.1.8) Where x and y represent respectively the input and output of a noisy communication channel, the conditional entropy $H_y(x)$ is termed the equivocation. It is a measure of ambiguity.

Equivocation

(5.1.9) The number of bits per second which a channel can transmit is termed its capacity. For the case (5.1.8) it is defined as the maximum of $(H(x) - H_y(x))$.

Capacity

(5.1.10) The ratio of the entropy of a source to the maximum value which it could have while using the same symbols is called its relative entropy.

Reletive entropy

(5.1.10.1) One minus the relative entropy is termed the redundancy.

Redundancy

(5.2) These considerations suggest a more economical method of communicating a representation.

(5.2.1) Instead of transmitting a physical representation of the representation itself, we may transmit a representation of the selection-process by which it may be identified in the ensemble of possible representations which is assumed to exist at the receiving end.

(5.2.2) A system whereby a representation is defined by a selection-process is termed a code-system.

Code system

(5.2.3) The corresponding representation of the selection-process transmitted is known as a code-signal.

Code-signal

(5.2.3.1) As a physical sequence the code-signal will itself have metrical and structural features as discussed in § 4, and will be definable by a vector in an information-space. But its structure need not have anything in common with that of the representation which it identifies.

(5.2.3.2) On the other hand the ordinary case of making physical representations could be thought of formally as a special case of coding, one-for-one.

(5.3) It follows that the result of an experiment, as well as a communication signal, could be analysed in terms of its selective information-content.

Selective information-content

Selective information-content) (5.3.1) This is a relative measure, depending on the number of distinct results which were regarded as equally probable by the observer. The result observed is thought of as specifying one of a number of possibilities already contemplated by the observer as forming an ensemble in defined proportions.

(5.3.2) The amount of selective information derived from the experiment can then be computed in the same way as for a message, (5.1).

(D) ALPHABETICAL INDEX OF TERMS USED IN INFORMATION THEORY AND RELATED COMMUNICATION THEORY. REFERENCES ARE TO PARAGRAPHS IN THE GLOSSARY. (STATISTICAL TERMS RELEVANT TO INFORMATION THEORY ARE LISTED SEPARATELY IN SECTION E).

Bandwidth: In general terms, the region of Fourier-space (q.v.) to which the output of an instrument is confined. In particular, the effective frequency-range (conjugate to a given coordinate) to which it responds.

(Binary digit
(Bit (5.1.3) Unit of selective information.

Capacity: (5.1.9) Number of bits transmissible per second.

Code-system: (5.2.2)

Code-signal: (5.2.3)

Conceptual unit: (4.2.11) The coordinate-interval in which one metron is acquired. Reciprocal of metron-density.

Ensemble: A set of possibilities each of which has a defined probability.

Entropy: (5.1.5) (a) In statistical mechanics, the weighted mean of the (negative) logarithm of the probabilities of members of an ensemble. (b) In thermodynamics that function of state of a body or system which increases by $\int_1^2 \Delta Q/T$ in a reversible process between two states 1 and 2, where ΔQ is the heat taken up by the body or system at temperature T. (Definitions (a) and (b) are equivalent).

(Conditional: (5.1.6)
(Relative: (5.1.10)

Equivocation: (5.1.8)

Fourier-space: The space whose dimensions represent variables which are Fourier-transforms of coordinates. (e.g. the frequency conjugate to the time coordinate).

Information: That which adds to representations (2.1).

(Metrical: (4.2) Specifying the number of elements of a pattern.
(Selective: (5.1.3) Specifying the unforeseeablesness of a pattern.
(Structural: (4.1) Specifying the number of independently variable features or degrees of freedom of a pattern

Information-matrix: A matrix by which the metrical information-content of an experiment is specified.

Information-space: The space in which independent groups of metrons are represented by orthogonal rays, and their metron-contents by the squares of distances along these rays. (4.3.1).

Information-vector: The vector whose components in information-space are the distances just mentioned.

Logon: Unit of structural information (q.v.)

Logon-capacity: (poss. syn: Logon-density) (4.1.4) Number of logons per unit of coordinate-space.

Logon-content: Syn: Structural Information-Content (4.1.3) Number of independently variable features.

Metron: (4.2.2) Elementary Unit of metrical information (q.v.).

Metron-capacity: syn: Metron-density (4.2.10) cf. logon-capacity.

Metron-content: Syn: Metrical Information-Content (4.2.4) Measures the amount of evidence to which a representation is equivalent.

Numerical Energy: (4.2.7.1) Ratio of (Energy)/(Noise Power per unit bandwidth). Analogous to metron-content.

Proper-scale: (4.2.6) A representational scale on which equal intervals are equiprobable.

Redundancy: (5.1.10.1) One minus relative entropy.

Representation: (1.1) A symbolic picture, model, statement, etc.

Scale-unit: (4.2.9) The minimum interval in terms of which a scale can definably or usefully be graduated.

(Metrical: (4.2.11) Undesirable equivalent of conceptual unit. Reciprocal of metron-density.

(Structural: (4.1.5) Reciprocal of logon-capacity.

(E) ALPHABETICAL LIST OF SOME STATISTICAL TERMS RELEVANT TO INFORMATION THEORY

Ensemble: A set of possibilities each of which has a defined probability.

Ergodic: Roughly speaking, statistically homogeneous. Applied to processes producing sequences, all samples of which are 'typical' if large enough.

Markoff process: A process in which a system changes from one to another of a finite number of possible states, according to a definable set of transition-probabilities (q.v.)

Stochastic process: A process producing a sequence of events determined only statistically.

Time-series: A sequence of numerical quantities distributed in time.

(Stationary: Such a sequence which is a member of a defined ensemble having a constant statistical character.

Transition-probability: The probability that a system in one state will go to another.

A HISTORY OF THE THEORY OF INFORMATION

by
E. Colin Cherry

(1) INTRODUCTION. LANGUAGES AND CODES

We who are gathered together here are drawn from many different fields of science: physics, physiology, mathematics, engineering.....; our common interest is the concept of information and the communication of information. In presenting this brief history, I shall attempt to trace a continuous thread through various fields of human activity from the earliest times to the present day, and attempt to show how this interest of ours has gradually developed, grown in scope and application, until it has become of such scientific importance as to attract an audience as wide and as distinguished as is attending this symposium.

Man's development and the growth of civilisations has depended in the main on progress in a few activities, one of the most important of which has been his abilities to receive, communicate and to record his knowledge. Communication essentially involves a language, a symbolism, whether this be a spoken dialect, an ancient stone inscription, a cryptogram, a Morse code signal, or a chain of numbers in binary digital form in a modern computing machine. It is interesting to observe that as the technical applications have increased in complexity, with the passage of time, the languages have increased in simplicity, until today we are considering the ultimate compression of information in the simplest possible forms. It is important to emphasise, at the start, that we are not concerned with what is being communicated; the meaning or the truth of a sentence is outside the scope of mathematical "information theory". The material which is to be communicated between a transmitting mind and a receiving mind must however be pre-arranged into an agreed language. We are not concerned with the 'higher' mental processes which have enabled Man to find a word for a concept, even less with the formation of a concept; we start only from the point when he already has a dictionary.

A detailed history of spoken and written languages would be irrelevant to our present subject, but nevertheless there are certain matters of interest which I should like to take as my starting-point. The early writings of Mediterranean civilizations were in picture, or "logographic" script; simple pictures were used to represent objects and also, by association, ideas, actions, names, and so on. Also, what is much more important, "phonetic" writing was developed, in which sounds were given symbols. With the passage of time, the pictures were reduced to more formal symbols, as determined by the difficulty of using a chisel, or a reed brush, while the phonetic writing simplified into a set of two or three dozen alphabetic letters, divided into consonants and vowels. [1, 2].

In Egyptian hieroglyphics we have a supreme example of what we now call "redundancy" in language and code; one of the difficulties in deciphering the Rosetta stone lay in the fact that a polysyllabic word might give each syllable not one symbol but a number of different ones in common use, in order that the word should be thoroughly understood[3]. (The effect, when literally transcribed into English, is one of stuttering). On the other hand the Semitic languages show an early recognition of redundancy. Ancient Hebrew script had no vowels; and modern Hebrew too, except in children's books. The vowels are merely indicated by dots and, depending upon what vowel sound the reader inserts, a word can sometimes assume different meanings. Many other ancient scripts show no vowels. Slavonic Russian went a step further in condensation; in religious texts, commonly used words were abbreviated to a few letters, in a manner similar to our present day use of Ampersand, e.g., etc., lb. and the increasing use of initials, T.R.E.; M.O.S.; Unesco and so on.

The Romans commonly wrote in abbreviated script, with little loss in meaning. We use the same in writing Latin script round our modern coins: "Dei Gratia, Brittania omnes Rex.....etc". But the Romans were responsible for a very big step in speeding up the recording of information. The freed slave Tyro invented shorthand, in about 63 B.C., in order to record, verbatim, the speeches of Cicero. This is not unlike modern shorthand in appearance (Fig. 1). It is probable that this was used for reporting the trials of the early Christian martyrs, and it is known to have been used in Europe until the early Middle Ages[4].

Related to the structure of language is the theory of cryptographs or ciphers, certain aspects of which are of interest in our present study, in connection with the problem of coding[24]. The act of ciphering is as old as the Scriptures, being of vital importance for military and diplomatic secrecy. The simple displaced alphabet code, known to every schoolboy, was most probably used by Julius Caesar[3]. One of the earliest signalling codes is described by Polybius, the Greek historian, 150 B.C. (Fig. 2) in which each letter was given two coordinate numbers one number representing a particular tablet and the other, one of five letters on the tablet, these numbers being signalled by torches held in the left and right hands[¥]. (Still used by Lloyds as late as 1780, as a commercial code[5]). There are many other historic uses of cipher; for example Samuel Pepys diary[4] was entirely ciphered "to secrete it from his servants and the World"; also certain of Roger Bacon's scripts have as yet resisted all attempts at deciphering. A particularly important cipher is one known as "Francis Bacon's Biliteral Code". This gentleman suggested the possibility of printing seemingly innocent lines of verse or prose, using two slightly different founts, which I will call 1 and 2. The order of the 1's and 2's was to be used for coding a secret message. Each letter of the alphabet was coded into 5 units, the founts 1 and 2 being used as these units[3]. Now, this code illustrates an important principle which seems to have been understood throughout history - information may be coded into a two - symbol code[¥¥]. There are numerous examples: bush telegraph signals of many Congo tribes use drum beats of notes with high and low pitch[ø]; long and short smoke signals are another case. Nowadays we have the Morse-code, in dots and dashes, and many similar codes.

Such two-symbol codes are the precursors of what we now call "binary coding", as used in pulse-coded telephony and high-speed digital computing machines. The transmitted information is coded into a series of electrical pulses and blank spaces, often referred to as a "yes-no" code. Now the ancient Celts invented a script which is of interest in this connection, known as the Ogam[4] script, found in Ireland and Scotland. Most scripts have developed into structures of complex letters, with curves and angles, difficult to chip in stone, but the Celts seem to have consciously invented this script, using the simplest symbol of all - a single chisel stroke - discovering that this was all that is necessary (Fig. 3). This script could be directly written in electrical pulses! Thus we could use positive pulses, negative pulses, double pulses and blank spaces. Our modern need is similar to that of the Celts, it is of great advantage for reasons of economy to be able to write using only one symbol - a pulse, or a chisel stroke.

¥ A. Belloc. "La Telegraphie Historique", 1894.

¥¥ Strictly, "logically communicable" information.

ø E.g. "The Drum Language of the Lokele Tribe"
J.F. Carrington. African Studies. 1944. Witwatersrand University Press

With the introduction of the famous dot-dash code by S.F.B. Morse, in 1832, a new factor was consciously brought in. This was the statistical aspect. Morse purposely designed his code so that the most commonly used letters were allocated the shorter symbols. (Fig. 4) It is of course true that in every language the most commonly used words have unconsciously become the shortest. The need for such a statistical view had been appreciated for a hundred years and, as time has progressed, this has assumed increasing importance. Certain types of message are of very common occurrence in telegraphy; for example certain commercial expressions and birthday greetings and these should be coded into very short symbols. By the year 1825 a number of such codes were in use. The modern view is that messages having a high probability of occurrence contain little information and that any mathematical definition we adopt for the expression "information" must conform with this idea, that the information conveyed by a symbol, a message or an observation, in a set of such events, must decrease according to their frequency of occurrence increasing. Dr. Shannon has emphasised the importance of the statistics of language, in his recent publications[24], and has referred to the relative frequencies of letters in a language, and of the digrams (or letter-pairs) such as ed, st, er, and of the trigrams ing, ter and so on, as occur in the English language. His estimate is that this language has a redundancy of over 50%, meaning that if instead of writing every letter (or its coded symbol) of an English sentence, we write suitable symbols representing the digrams, trigrams, etc., the resulting compression would be about two to one. If such a perfect code be used, any mistake which might occur in transmission of a message cannot be corrected by guessing. Language statistics[3] have been of essential interest for centuries, for the purpose of deciphering secret codes and cryptograms. The first table of letter frequencies to be published was probably that of Sicco Simonetta of Milan, in the year 1380; another, used by Porta in 1658 used digrams also.

Modern mathematical symbolism illustrates a language possessing a high degree of compression of information. The Greeks had largely been limited to geometry, algebra eluding them because of the lack of a symbolism. Descarte's application of formulae to geometry and, even more important, Leibnitz' great emphasis on symbolism are two outstanding developments in the compression of mathematical information. The importance of symbolism is indeed prominent throughout the modern evolution of mathematics, as its generalisations have increased; Russell and Whitehead's treatment of the bases of mathematics (1910) as a generalisation of ordinary logic was written almost entirely without words[6]. During the last century, the idea has emerged of mathematics as the "syntax of all possible languages" and as the language of logic. In particular Peano[X] invented the symbolism used for symbolic logic.

Leibnitz not only considered mathematical symbolism as a universal "language of logic", but he was a great advocate of language reform. Already, in 1629, Descartes had considered the possibility of creating an artificial, universal, language, realising that the various languages of the world were, by virtue of their complex evolution, utterly illogical, difficult and not universal. However, his dreams did not materialize until 1661, when the Scotsman George Dalgarno published his "Ars Signorum". All knowledge was firstly to be grouped into 17 sections, under headings such as "politics", "natural objects" etc., and each heading represented by a consonant; each section was then to be divided into subsections, represented by a vowel, then into sub-sub-sections and so on, consonants and vowels alternating[2]. Every word, always pronouncable, thus denoted an object or an idea by a sequence of letters representing selections, from the prearranged sections, sub-sections etc; this notion of selection is very relevant to the modern theory of information and will be discussed more fully later. It will be recognised that such a language structure has some resemblance to Chinese written characters and that it has equally the lack of flexibility to incorporate new

---

X See C.I. Lewis "Survey of Symbolic Logic"

ideas and knowledge as time progresses. Leibnitz was aware of this serious restriction with the Dalgarno language and his own proposals were based on a comparative analysis of natural languages, in order to construct a rational grammar, by the rejection of all redundant devices such as gender, the removal of irregularities, the simplifying of conjugation and so on.

This work of Leibnitz, on the construction of a rational grammar, has no real connection with information theory, since it is essentially concerned with the meaning of a sentence; however the ideas of Descartes and Dalgarno, to a limited extent, have. Information theory starts when an idea has already been expressed as a sentence or a succession of letters or symbols, each successive symbol representing a selection out of a pre-arranged language or code, possessing a certain statistical structure. In his recent work, Dr. Shannon[24] has illustrated the idea of the building-up of a "message" as a stochastic process, that is as a series of words, each one being chosen on a statistical basis, depending on the one, two, three..... words immediately preceding. That such sequencies of words can bear some resemblance to an English text merely illustrates the accuracy of the statistical tables used, although no meaning is conveyed by the resulting "message". One is reminded here of the old fable concerning a hundred monkeys tapping on a hundred typewriters producing every book in the world, given infinite time. In this case the letter sequences are presumably utterly random and the question reduces to what is now known as the problem of noise, a sequence containing no information whatever.

The origin of this tale is interesting, since it is most certainly from Jonathan Swifts' "Gullivers Travels",# Gulliver has paid a visit to the Academy of Lagado, and amongst a series of abusive descriptions of imaginary research programmes, describes that of the Professor of Speculative Learning: "The professor observed me looking upon a frame which took up a great part of the room; he said that by this contrivance the most ignorant person may write in philosophy, poetry and politics. This frame carried many pieces of wood, linked by wires. On these were written all the words of their language, without any order. The pupils took each of them hold of a handle, of which there were forty fixed round the frame and, giving them a sudden turn, the disposition of the words was entirely changed. Six hours a day the young students were engaged on this labour; the professor showed me several volumes already collected, to give to the world a complete body of all the arts and sciences. He assured me that he had made the strictest computation of the general proportion between the numbers of particles, nouns and verbs.....". One wonders what would be Dean Swifts' reactions, could he attend the present Symposium!

(2) SIGNAL COMMUNICATION THEORY

Perhaps the most important technical developments which have assisted in the birth of our subject of "information theory" are those of telephony and telegraphy. With their introduction, the idea of speed of transmission arose and when their economic value was fully realised, the problems of compressing signals exercised many minds, leading eventually to the concept of "quantity of information" and to theories on times and speed of signalling. In the year 1267 Roger Bacon[4] suggested that what he called "a certain sympathetic needle" (i.e. Lodestone) might be used for distant communication. Porta and Gilbert, in the 16th century, wrote about "the sympathetic telegraph" and in the year 1746 Watson, in England, sent electric signals over nearly 2 miles of wire. In 1753 an anonymous worker used one wire for each letter of the alphabet, but in 1787 Lomond used one wire-pair and some code. Gauss and Weber invented a 5-unit code in 1833, later to be named after Baudot, in honour of his life's work in automatic telegraphy. The introduction of carrier waves, during the first World War,

In 1726. "The Voyage to Laputa" Chapter 5 (the version above has been condensed)

was made practicable by G.A. Campbell's invention of the wave filter and the method of frequency-division multiplex rapidly developed, some of the earliest work being carried out by Colpitts and Blackwell (1921). This principle of allocating simultaneous signals into "frequency-bands" has been the mainstay of electrical communications and has remained unchallenged until the recent War.

Related techniques which have greatly urged the development of general communication theory are those of telephony and television. Alexander Graham Bell's invention of the telephone in 1876 has particular significance in relation to our present physiological interests, to which I shall refer later; otherwise it is, from our present point of view, purely a technological development, setting up problems similar to those of telegraphy. However, early in the history of television, 1925-27, the very great bandwidth required for detailed "instantaneous" picture transmission was appreciated and this was brought to a head with the introduction of the techniques of cathode-ray-tubes, mosaic cameras and other electronic equipment rendering high-definition practicable. Great masses of information had now to be read off at high speed, at the camera end, transmitted and reassembled at the receiver. Now that the human eye had been brought into the communication link, with its ability to recognise shape, the problem of phase-distortion became all important, although Sallie Pero Meade had been concerned with phase troubles in long distance telegraphy, in 1928. Furthermore, other major theoretical studies were forced by the great band-widths required for television; in particular the noise problem, the transient response problem, and the problem of amplifying wide-bands of energy. Noise is of particular interest to our subject, as we are all aware; it is the ultimate limiter of the transmission of information. The subject of noise is itself a vast one and cannot be treated in this short history, save only to mention the names of the chief pioneers, A. Einstein (1905)[49], Schottky (1918)[50], Johnson[52] and Nyquist (1928)[53].

But to return for a moment to the first World War: wireless had been developed from the laboratory stage to a practical proposition largely due to the field work of Marconi and the early encouragement of the British Post Office, at the turn of the century[4]. Sidebands were soon discovered by many people and serious but fruitless controversies arose as to whether they did or did not exist. The advantages of reducing the bandwidth required for the transmission of a telephony signal were appreciated, but the early theories of modulation were very vague and lacked mathematical support. In 1922 John Carson[9], of the A.T.T. Company, clarified the situation, with a paper showing that the use of frequency modulation, as opposed to amplitude modulation, did not compress a signal into a narrower band. He also made the important suggestion that all such schemes "are believed to involve a fundamental fallacy", a fact we now know well. At this time it was well known that only one sideband needs to be used, since both contain the same information[¥]. Curiously enough, although it was recognised, very early on, that carrier telephony requires the same bandwidth as the spoken word, nobody seems to have stated clearly that carrier telegraphy requires a finite bandwidth until long after such systems had been in use. At that time, the inventors of wireless telegraphy naively imagined that the "frequency" was "on" for a time and "off" for a time. The priority again probably belongs to Carson.

In 1924, Nyquist[10] in the United States and Küpfmüller[11] in Germany simultaneously stated the law that, in order to transmit telegraph signals at a certain rate, a definite bandwidth is required, a law which was expressed more generally by Hartley[12] later, in 1928. This work of Hartley's has a very modern ring about it; he defined information as the successive selection of symbols or words, rejecting all "meaning" as a mere psychological factor, and showed that a message of N symbols chosen from an alphabet or code of S symbols has S^N possibilities and that the "quantity of information" H, was most

---

¥ Carson, 1915, patents. Espenscheid, 1922, demonstration.

reasonable defined as the logarithm, that is H = Nlog S. Hartley also showed that in order to transmit a given "quantity of information" a definite product, bandwidth X time, is required. We shall later be considering the more modern aspects of this theory of Hartley's, which may be regarded as the genesis of modern theory of the communication of information.

After Carson, the next major contribution to the theory of frequency modulation is due to Balth van der Pol in 1930 who used the concept of "instantaneous frequency" as originally defined by Helmholz[X] as rate of change of a phase angle.

In 1936, Armstrong published an account of the first practical frequency-modulation system[14] which appeared to refute Carson's results. The new principle was used of limiting a carrier to constant amplitude, by which means the strongest of a number of carrier waves, received simultaneously, might be separated out (the "capture effect"). The advantage was claimed for the system that a number of stations might be sited close together, and the strongest received at a given point without interference from the others. Also the system offered advantages with regard to signal-to-noise ration, a point not considered by Carson in 1922.

All the early modulation theories took as a basic signal the continuous sine wave or, at the best, a continued periodic waveform. Such applications of Fourier analysis give "frequency descriptions" of signals and are essentially timeless. The reverse description of a signal, as a function of time, falls into the opposite extreme, as if the values of the signal at two consecutive instants are independent. Practical signals, whether speech or coded symbols, are of finite duration and at the same time must, to be reasonable, be considered to occupy a certain bandwidth. The longer the signal time-element $\triangle$ t, the narrower the frequency-band $\triangle$ f or, as we may say, the more certain is its frequency.

Gabor[15] took up this concept of uncertainty in 1946, and associated the uncertainty of signal time and frequency, $\triangle$ t. $\triangle$ f $\simeq$ 1, by analogy, with the Heisenberg uncertainty relation of wave mechanics. In this, he is most careful to point out that he was not attempting to explain communications in terms of quantum theory, but was merely using some of the mathematical apparatus. Gabor points out that our physical perception of sound is simultaneously one of time and frequency, and that a method of representation may be used which corresponds more nearly to our acoustical sensations than do either the pure frequency-description or time-description. The basic signals on which such a representation is based, must be finite in both time and frequency bandwidth. Using reasoning closely related to that of Hartley[12], Gabor shows that there must be only two independent data expressed by these basic signals, per unit of time X frequency product (i.e. 2f.t. data)†. There are many choices for the basic signals, but one of particular interest uses sine-waves modulated by a Gaussian probability function. These are distinguished by their property that their effective bandwidth x time-duration product is the smallest out of all possible signals and hence they overlap as little as possible. Such a signal is regarded as a "unit of information" and is called by Gabor a logon. Further, their Fourier transforms have the same mathematical law; hence the representation (which is made as a kind of matrix) is symmetrical in frequency and

---

X Helmholz, "Die Lehre von den Tonempfindungen" 1862.

† As implied by Küpfmüller[11] and Nyquist[10] in 1924. These data need not necessarily be thought of as Fourier coefficients but, for example, as in pulse-modulation, samples of a waveform at intervals 1/2f for a time t. Such samples may be shown to define the waveform.

time. Also they have the advantage that the notions of amplitude and of phase can be applied to these signals as well as, and with more physical justification than, to continuous sine-waves. In this work, Gabor operates with complex signals, which have no negative frequencies; more recently the theory of such complex signals has been extended by J.A. Ville, in France[16]. Such logons relate to a given channel and not to any particular signal; such units are now called units of "structural information" and define the units of which any transmitted signal may be considered to be composed.

By this time in the history of communication it had been realised for several years that in order to obtain more economical transmission of speech signals, in view of this bandwidth-time law, something drastic must be done to the speech signals themselves, without seriously impairing their intelligibility. These considerations led to what is known as the "Vocoder", an instrument for analysing, and subsequently synthesising, speech. It is fair to say that this arose out of a study of the human voice and the composition of speech, which is itself of early origin; for example Graham Bell and his father had studied speech production and the operation of the ear, while more recently there has been the work of Sir Richard Paget[17] and of Harvey Fletcher[18]. In year 1939 Homer Dudley[19, 20] demonstrated the "Voder" at the Worlds Fair, New York. This instrument produced artificial voice sounds, controlled by the pressing of keys and could be made to "speak" when manually controlled by a trained operator. In 1936 Dudley had demonstrated the more important "Vocoder": this apparatus is essentially a means for automatically analysing speech and reconstituting or "synthesising" it. The British Post Office also started, at about this date, proceeding on an independent programme of development, largely due to Halsey and Swaffield[21]. In simple terms it may be said that the human voice employs two basic tones: those produced by the larynx operation, called "voiced sounds", as in ar, mm, ooh, and a hissing or breath sound, for which the larynx is inoperative, as in ss, h, p, etc. Speech contains much which is redundant to information or intelligence and which is therefore wasteful of bandwidth; thus it is unnecessary to transmit the actual voice tones, but only their fluctuations. At the transmitter these fluctuations of the voice tones are analysed and sent over a narrow bandwidth. At the same time another signal is sent to indicate the fundamental larynx pitch, or if absent, the hissing, breath, tone. At the receiver, these signals are made to modulate and control artificial locally produced tones from a relaxation oscillator or a hiss generator, thus reconstituting the spoken words.

The minimum overall bandwidth required is (at least on paper) about 275 c/s, a compression of about 10/1; such a method may be regarded as a voice operated coder, which automatically codes and decodes the voice. Two decades earlier Küpfmüller had suggested that a compression ratio of 40/1 in the magnitude of $\Delta f. \Delta t.$ can be achieved, for the transmission of a single letter, if a telegraph code be used rather than the human voice. Another method to reduce the bandwidth of the signal, called "frequency compression", has been described by Gabor[22]. In the transmitter a record of the speech is scanned repeatedly by pick-ups themselves running, but with a speed different from that of the record. Thus a kind of Doppler effect is produced, reducing the bandwidth of the transmitted signal, which in the receiver is expanded to its original width by a similar process. It is of course impossible to reduce all frequencies in the same ratio, since this would imply stretching the time scale; what the apparatus does is rather to reduce the acoustical frequencies, leaving the syllabic periods unchanged. It has been demonstrated that four-fold compression is possible with hardly any, and sixfold compression with slight, loss of intelligibility, but there is some loss of quality. This can be avoided, according to Gabor's theory, by matching the scanning frequency continuously to the voice pitch, but this has yet to be demonstrated experimentally. We have so far considered, in the main, the frequency aspect of the transmission of signals, that is, questions of bandwidth; of frequency spectra of signals; of amplitude, phase or frequency modulation and so on. This frequency aspect absorbed most attention during the very rapid

developments of the 1920's and early 30's. In the last few years, however, it is the time aspect of which we hear so much; of pulse modulation; of pulse code modulation and of time division multiplex. Referring back into history, the earliest suggestion for the simultaneous transmission of two messages, over one line without frequency-separation, seems to have come from Edison and Heaviside who introduced the "duplex" and "quadruplex" systems† in 1873-4 (Fig.5). With this system one message, sent in Morse code, was read at the receiving end by a polarized relay; the second message was transmitted as an amplitude modulation of the first signal and was read by an unpolarized relay, the first relay, the first message merely acting as a carrier wave and so ignored by this unpolarized relay. The important principle was employed here that two messages can be sent simultaneously, over the same bandwidth that is required for one, if the power is increased. Although not explicitely stated in this form in his paper, Hartley[12] has implied that the quantity of information which can be transmitted is proportional to the product: f. t. Nlog S, where S is defined as the number of "distinguishable amplitude levels". Hartley has considered messages consisting of discrete symbols, for example letters or Morse code, and also messages consisting of continuous waveforms, such as speech and music. He observes that the latter signals do not contain infinite information since "the sender is unable to control the waveform with complete accuracy". He approximates the waveform by a series of steps, each one representing a selection of an amplitude level. Such a representation is nowadays referred to as amplitude quantisation of the waveform. For example, consider a waveform to be traced out on a rectangular grid (Fig. 6), the horizontal mesh-width representing units of time and the vertical the "smallest distinguishable" amplitude change; in practice this smallest step may be taken to equal the noise level, n. Then the quantity of information transmitted over a channel may be shown[23, 24] to be proportional to

$$\text{f. t. log } \left(1 + \frac{a}{n}\right) \text{ where a} = \text{the maximum signal amplitude.}$$

The total transmitted quantity of information may be held constant, but the magnitudes f, t, a, changed; thus bandwidth or time may be traded for signal power. This principle has been given practical embodiment in various systems of pulse-modulation[25] and time-division multiplex. Although a number of such systems had appeared in Patent form in the 1920's, their application was mainly to improvement of transmitter efficiency. Their development was delayed until a few years before the recent War, partly owing to the lack of suitable pulse and micro-wave techniques. Particular mentioned should be made of the work of Reeves and Deloraine[x], who patented time-division multiplex systems[26] in 1936. In such systems the waveform is not transmitted in its entirety, but is "sampled" at suitable intervals, and this information is transmitted in the form of pulses, suitably modulated in amplitude, width, number or time-position. Reeves[ø] also proposed another system, which uses directly the idea of amplitude quantisation, as envisaged by Hartley; the waveform is automatically restricted, at any instant, to one of a number of fixed levels (as illustrated in Fig. 6) before being sampled and transmitted as a pulse signal. This assimilation of telephony into telegraphy becomes even more complete if the quantised pulses are coded[26]. Thus the information which must be transmitted is given by the number of quanta in each amplitude sample and this number, having one of S possible values, may be coded into a train of pulses in, for example, binary (yes-no) code.

---

x Brit. Patents 509,820 and 521,139 (U.S. Patent 2,262,838)
Brit. Patent 511,222 (U.S. Patent 2,266,401;
French Patent 833,929 and addition 49,159)

ø Brit. Patent 535,860 (U.S. Patent 2,272,070. French Patent 852,183)

A waveform quantised in amplitude and time, as in Fig. 6, can have S^N possible "states"; regarding these states as an alphabet of all the possible waveforms, Hartley's law gives the information carried by one waveform as $H = K. N \log S$, which is finite. Since Hartley's time this definition of information as a selection of a symbol has been generally accepted, variously interpreted and gradually crystallized into an exact mathematical definition.

Any quantitative description of the information in a message must be given in statistical terms; the information conveyed by a symbol must decrease as its probability increases. Attaching probabilities to the various symbols P_1 P_2 ..... P_i ..... in a message, or to the various "states" of a waveform, Hartley's law may be re-interpreted so as to define the total information in a sequence of n symbols as, on an average:

$$H_n = - \sum_n P_i . \log P_i$$

an expression which has been evolved, in various ways by several different authors(X), in particular Shannon[24], Wiener[37], Fano[27] during the last few years. The unit of information H, in this selective sense, has been called the "bit" (from Binary Digit).

This expression for the information is similar to that for the entropy* of a system with states of probabilities P_1 P_2 ..... P_i ... , using the term in the Boltzmann statistical sense. Probably the first detailed discussion of the identity between information and entropy was made by Szilard[48] as early as 1929 who, in a discussion on the problem of "Maxwell's Demon", pointed out that the entropy lost by the gas, due to the separation of the high and low energy particles was equal to the information gained by the Demon and passed on to the observer of this "experiment". In his recent publications, our distinguished guest Dr. Shannon has forged this theory of communication of information[24] into a coherent theory, using the Boltzmann statistical definition of entropy as a basis. We have already referred to his treatment of a message, or a succession of symbols, as a process (called a stochastic process) governed by probabilities and transition probabilities, and to the idea of redundancy in a language or a code#. Zero redundancy corresponds to maximum entropy, since then nothing is known a priori about the message; the information content is perfectly compressed. If the transmission channel has noise, the message may in theory be so encoded as to achieve maximum entropy in this channel - thereby transmitting information at the greatest possible rate, consistent with the limitations of the capacity of the particular channel. If noise is present, an encoding system exists, which introduces just sufficient redundancy to overcome this noise, giving as low a frequency of errors as desired (we are all familiar with the idea of having to repeat words or sentences on a noisy telephone). These theorems and many others have been expounded by Shannon, in precise mathematical form and are of the most general application.

The related problem of designing the apparatus comprising the communications channel, on a statistical basis, was considered earlier by Wiener/, who dealt with the design of a filtering system to be optimal in combatting noise, showing that an ideal response characteristic exists which

---

(X) Including W.S. Percival, in Britain, in unpublished work (1939).

* The relation to the entropy concept of statistical mechanics is not exact. It has therefore been decided to use the term "selective entropy"; this term has been included in the "glossary".

A.A. Markoff, early in this century, was probably the first to treat written messages as "chains" governed by probabilities and transition probabilities.

/ N. Wiener "The Interpolation, Extrapolationaand Smoothing of Stationary Time Series" John Wiley, 1949.

minimises the mean square error between waveforms of the input signals and the output signals plus noise. A time delay is the price paid; the longer the time delay, the less is the chance of error. A related problem considered by Wiener is that of prediction. He shows that a predictor circuit may be designed which reproduces the incoming waveforms with minimum mean square error, in advance of their arrival by a short time; in this case the longer this time advance, the greater is the chance of error. In both problems a certain delay period is required as the price of a solution; it is the price of the increased information. The choice of a mean square error is an arbitrary one and may possibly be open to question.

(3) "BRAINS"; REAL AND ARTIFICIAL

Progress in computing methods and the development of computing machines, in modern times, has been forced by the increase of statistical analyses in science and business, calling for rapid treatment of masses of numerical data. The development of the most recent digital computing machines, such as the Eniac[28] in U.S.A. and the Edsac and A.C.E.[29] in Britain, the so-called "electronic brains" primarily for application to problems in mathematical physics and in pure mathematics, raise more complex problems in programming, that is, the breaking down of mathematical operations into the most elementary steps and the logical feeding of these steps into the machine together with a priori data referring to the particular calculationϕ. The surprising thing is that, once the mathematical processes have so been broken down, both these fundamental steps and the actions required of the machine are few in number and elementary in principle; such simple processes as adding, subtracting, "moving-up one" etc.

Descartes and more particularly Leibnitz, had visions of computing machines - "reasoning machines" as they were considered.$\not$ But the lack of technique prevented practical construction until Charles Babbage, while Professor of Mathematics at Cambridge between 1829 and 1838, commenced construction of two different kinds of automatic digital computing machines (unfortunately never completed), one of which bears considerable resemblance, in its basic structure, to our modern machines.[29,30,31] This "analytical engine" possessed three component parts: a "store" for data or for intermediate results of a calculation, which could be read off as desired, a "mill" for performing arithmetical operations and an unnamed unit (actually a "controller") for selecting the correct data from the store, for selecting the correct operation for the "mill" and for returning the result to the store.

ϕ The punched card system, for storing numbers and subsequently entering them into a machine, was first conceived by Herman Hollerith, in 1889; the scheme is similar to that used for weaving of patterns in cloth, by the Jacquard loom, which is itself essentially a problem in coding.

$\not$ Pascal constructed an adding machine (using numbered wheels) in 1642; Leibnitz built a digital multiplying machine in 1694. The modern desk computing machines, such as the Marchant type, have originated from these. The "reasoning machines", to which later reference will be made, were visualized as dealing with problems of logic and not merely for arithmetic computation[4].

Over a 100 years later, the first successful mechanical digital computor was operated, the Harvard Mark I calculator, developed by Aiken[28], using the fundamental principles envisaged by Babbage.[32] The first electronic machine was the Eniac (Eckert and Mauchly), with its inherent advantages of speed. Although the choice of a suitable scale or radix is quite arbitrary for the representation of numbers in a machine, there are great advantages offered by a binary scale, using only the numbers 0 and 1, since the physical elements used for representing numbers in a machine usually have two mutually exclusive states. For example, a switch, a relay or a valve can either be on or off. Such a method of representing numerical information is another example of the age-old coding principle, - information can be represented by a two-symbol code. The operation of a computing machine is thus of the same nature as that of any electrical communication channel; information is supplied from a "source", suitably coded, transmitted, operated on in various ways, and passed to the output. From the information theory point-of-view there are however, certain differences. Firstly a computing machine is usually "noiseless" in that it cannot be allowed to make a single mistake[¥], since this mistake would render all subsequent calculations invalid; it may however possess a limiting accuracy, set by the limited digital capacity. Secondly, the machine comprises many individual communication channels[ø]. Thirdly the questions of the language statistics and coding such as arise in electrical communications, are replaced by problems of programming, or breaking down a calculation into elementary steps and feeding these as instructions together with other numerical data into the computing machine in the correct sequence, as the "input information". It is the automatic feeding-in of the sequence of instructions, which distinguishes these modern machines from the manually operated desk types and especially the facility of changing the sequence according to criteria evaluated during the course of calculation.

A suitable diagrammatic notation, for representing schematically the functional operation of these binary digital machines was suggested by von Neumann, and later modified by Turing, being adapted from the notation proposed by Pitts and McCulloch for expressing the relations between parts of the nervous system. This pair of workers[34] had applied the methods of mathematical logic to the study of the union of nerve fibres, by synapses, into networks while Shannon, in his doctorate thesis, had applied Boolean algebra to electric circuit switching problems. Now the nervous system may be thought of, crudely speaking, as a highly complex network carrying pulse-signals, working on an on-or-off basis; a neuron itself is thought to be a unit, like a switch or valve which is either on or off. The rise of a common notation for expressing the actions of the nervous system and of binary computing machines, at least recognises a crude analogy between them.

It is this analogy which has become greatly extended and has today become a considerable scientific interest. One is naturally led to extend the analogy to thought processes and to the possible design of "reasoning machines" thus accomplishing the dream of Leibnitz. Just as arithmetic has led to the design of computing machines, so we may perhaps infer that symbolic logic may lead to the evolution of "reasoning-machines" and the mechanization of thought processes.

ø Either the serial system may be used, in which all the digits of a binary number are transmitted in time sequence, or the parallel system, in which they are sent simultaneously over different channels; this is analogous to the time-bandwidth alternative, as in electrical communications channels.

¥ Self-checking procedures are now a recognised essential part of high speed digital computor operation; see Ref.33 (also for a further bibliography on this subject). Errors can be detected, and possibly corrected, either by duplication of the apparatus or, more economically, by the introduction of additional redundant digits. These digits enable errors to be located, the price being a slightly slower operation of the machine.

Such possibilities, together with the success already achieved by the automatic computing machine has caught the popular imagination during the last few years. On the whole, the approach of the scientist to this field has been wise and cautious, but its news value has somewhat naturally led to exaggerations in the lay press; thus, phrases such as "electronic brain", "machines for writing sonnets" are now commonly heard, and may have led to some scepticism in certain scientific circles. However, practical accomplishments are now forcing one to the conclusion either that "mechanised thinking" is possible, or that we must restrict our concept of "thinking". Much analysis of "mechanised thinking" has been promoted by the theory of (intellectual) games[35]. Machines which play "noughts and crosses" and, incidentally, will always win or draw /, are comparatively simple; the existence of such machines does not often cause surprise, because this game is considered determinate - presumably owing to the very limited numbers of moves - whereas chess or card games give one the feeling that "judgement" is involved. However, Shannon has recently dealt with the problem of programming a computor for playing chess[36], concluding that a machine is constructible, in principle, which can play perfect chess, but that owing to the great number of possible moves, this would be impracticable. Nevertheless one could be made which would give its opponent a very good game. Turing had also previously referred to such possibilities in unpublished work. There are two important points, which are emphasised by most writers concerned with "mechanised thinking". Firstly, the machine acts on instructions given to it by its designer; as an illustration, Shannon observes that at any stage in a game of chess, played against his machine, the next move which the machine will make is calculable by its designer, or by one who understands its programming.† Again, every single step in a calculation, carried out by a digital computor, could well be done by a human - the machine is merely far quicker. The second point of emphasis is the importance of the programming, rather than the machine in the metal. As Norbert Wiener has been most careful to stress[37], it is not the machine which is mechanistically analogous to the brain, but rather the operation of the machine plus the instructions fed to it. If a machine can ever be said to learn by its own mistakes, and improve its operation[33], it can only do this, as Shannon emphasises in connection with chess, by improving its programming.

But let us digress for a moment. During the years immediately preceding the recent war, it was becoming apparent that there existed similarities of ideas, basic concepts and methods, which formed a "no-mans-land" between certain various specialised branches of science. For two hundred years no single man has been able to compass the whole of science; the intensity of specialization has steadily increased and necessarily been accompanied by much duplication of work. Nowadays, to venture out of one's own recognised domain of research is to invite an accusation of dilettantism. The lead was taken by Norbert Wiener[37] who, with Rosenbleuth, used an old word "Cybernétics"¥ to name this no-mans-land. The needs of the recent war brought matters to a head, with the urgency of developing not only high speed computing machines, but

---

/ For example, work has been carried out at the National Physical Laboratory, England. See article by D.W. Davies, in "Science News" (Penguin Ltd.).

† The operation of machines is deterministic, unless an element is included which permits a random choice to be made. For example see D.M. Mackay "the Electronic Brain and its Philosophical Implications" The Christian Graduate, Sept. 1949, and A. Turing, "Computing Machinery and Intelligence". Mind. Oct. 1950, Vol.59, No.236, p. 433.

¥ Greek κυβερνητης meaning "steersman". The word "Cybernétique" was first coined by Ampère, to mean "the science of Government" ("Essai sur la Philosophie des Sciences", Paris, 1834).

automatic predictors, automatic gun laying mechanisms and other automatic following or "self-controlling" systems and to these two scientists should be given credit for calling attention to the need for a general study to include, not only these automatic mechanisms, but certain aspects of physiology, the central nervous system and the operation of the brain, and even certain problems in economics concerning the theory of booms and slumps. The common thread here, linking these topics, whether mechanistic, biological or mathematical, is the idea of the communication of information and the production of a self-stabilising control action. Apart from a study of the mechanical governor by Maxwell, in 1868, the first mathematical treatment of the stabilisation of a dynamic system by feeding back information from the output or "receiver" end to the input or "transmitter end" was made by H.S. Black, in a study of electrical feedback amplifiers[38], in 1934, later developed, largely due to the efforts of Nyquist [39] and of Bode [40], into an exact mathematical method and a system of design. The extension of the principles to electro-mechanical or to purely mechanical systems was a logical and natural one and the design of automatic following systems, such as those for anti-aircraft guns, for automatic pilots in aircraft, etc., need no longer proceed entirely on a trial and error basis.

For these automatic control systems, the word "Servo-mechanism" has been coined. The existence of numerous controls in the body accounts partly for a common interest with physiology. For example, there is homeostasis or the involuntary control of body temperature, of heart rate, blood pressure and other essentials for life, while voluntary control is involved in muscular actions, such as those required for walking along a narrow plank; the simplest movement of a limb may involve multiple feedback actions. If a stabilised Servo-mechanism has its feedback path open-circuited, so that the magnitude of its error cannot be measured at the input end and so automatically corrected, it is liable to violent oscillation; an analogous state of affairs in the human body has been mentioned by Wiener[37], called 'ataxia', corresponding to a nervous disorder which affects the control of muscular actions. The analogies in physiology are countless; Wiener goes even so far, in developing the analogy between the operations of a digital computing machine and of the brain and central nervous system, as to compare certain mental, functional, disorders (the laymans "nervous breakdowns") to the breakdown of the machine when overloaded with an excess of input instructions as, for example, the storage or "memory circuits" cannot store enough information to be able to tackle the situation. Note again, the emphasis is on the operation of the machine together with its instructions; no material damage may have occurred.

One is led instinctively to ask whether such analogies are not modern examples of a kind of "animism" / though these analogies do not imply any attempt to "explain" life on a mechanistic basis or to explain the body as a machine in the sense of Descartes who observed that the action of the body, apart from the guidance of the will "does not appear at all strange to those who are acquainted with the variety of movements performed by the different automata, or moving machines fabricated by human industry.....Such persons will look upon this body as a machine made by the hand of God".

Early invention was greatly hampered by an inability to dissociate mechanical structure from animal form. The invention of the wheel was one outstanding early effort of such dissociation. The great spurt in invention which began in the sixteenth century, rested on the gradual dissociation of the machine from the animal form, with a consequential improvement in method and performance, until machines eventually became completely functional. The development of machines had a converse effect, and the body came to be regarded as nothing but a complex mechanism: the eyes as lenses, the arms and legs as levers, the lungs as bellows, and so on. Julien de la Mettrie, in about 1740

/ For example see Lewis Mumford "Technics and Civilisation".

wrote, "(he thought) that the faculty of thinking was only a necessary result of the organisation of the human machine", a materialist view which so greatly disturbed the 'vitalists' of the time[6]. Since "animistic thinking" has been recognised as such, by inventors and scientists, its dangers are largely removed and turned to advantage, though amongst laymen it sub-consciously exists today (the very use of the expression "electronic brain"). Physics and biology have gone hand in hand; for example Harvey's discovery[6] of the circulation of the blood (1616) owes much to the work being carried out on air-pumps in which Harvey was interested. In more modern times, Helmholz attempted to unify physics physiology and aesthetics, in his studies of music and hearing. Electrical communications owes a debt to physiology; thus Graham Bell[4], the son of A.M. Bell who was an authority on phonetics and defective speech and who invented a system of "visible speech", became Professor of Physiology at Boston in 1873 and invented the first telephone after constructing a rubber model of the tongue and soft parts of the throat. The telephone receiver also was modelled on the structure of the human ear. At the present day, comparative studies of the central nervous system and of the operation of automatic computing machines will undoubtedly be to mutual advantage.†

Although reflex response[6] had been observed in the 16th century and the essential function of the spinal cord in 1751, the relation between function and structure remained elusive until 1800. In 1861 Broca fixed on the area of the cortex concerned with speech while Thomas Young, in 1792 had settled on that part associated with the eye[6]. The "on-or-off" action of nerve cells was first discovered by Bowditch in 1871, but the neuron theory, that the entire nervous system consists of cells and their outgrowths has only been developed during the present century[6]. That the intensity of nerve signals depends on the frequency of nervous impulses was observed by Keith Lucas, in 1909, work which was subsequently carried to an extreme elegance with the assistance of modern amplifier and oscillograph technique by Professor Adrian[6], in the late 1920's.

It is most certain that further studies in physiology will lead to new developments in electrical techniques which in turn may reflect back; new theories and generalities may emerge, leading to greater understanding of machine capabilities. The study of the behaviour of these machines and methods of their control or programming may cast new light on logic, as Turing has suggested. Already there are signs of far reaching developments.

The philosopher John Locke considered the content of the mind to be made up of ideas, not stored statically like books on a shelf, but by some inner dynamical process becoming associated in groups according to principles of "similarity", "contiguity" or "cause and effect". The word idea meant "anything which occupied the mind" or "any object of the understanding".‡ The first major experimental work, inherently implying a physiological basis for operation of the mind, was not carried out until Pavlov, starting about 1898, studied "patterns of behaviour" of animals. He produced salivation in a dog by showing it certain objects which had already been associated in the dog's mind, by previous repetition, with food - the "conditioned reflex". It seems likely that it must have been understood, in a descriptive way that an action was taking place here that we now called "feedback". Wiener[37] has taken the view that conditioned

---

† As emphasised by Prof. J.Z. Young in the B.B.C. Reith Lectures, 1950.

‡ "Essay Concerning Human Understanding" (1690).

reflexes enter into the field of "Cybernetics"; that, to give the extremes, the encouragement of actions which lead to pleasure in the body and the inhibition of those which lead to pain may possibly be regarded as feedback actions, suggesting inter-connections between different parts of the nervous system. Further, he observes that a conditioned reflex is a "learning mechanism" and that "there is nothing in the nature of the computing machine which forbids it to show conditioned reflexes". Again, 'machine' here includes its instructions.

Pavlov's work was continued during the first decade of this century and developed into the subject[4] of "Behaviourism"; the British psychologist Lloyd Morgan founded the American school of animal psychology, which has flourished largely under the guidance of J.B. Watson, to study learning and other aspects of behaviour of animals. Such psychological studies are closely accompanied by neurological work, so that gradually the parameters of mental experience, as originally envisaged by Locke, are becoming more accurately defined. Experimental work on the crudely parallel "behaviourism" of machines is at present being conducted, in particular in Britain by Dr. Grey Walter[41]. The inaccessibility and complexity of the central nervous system and of the brain render direct analysis overwhelmingly difficult; the brain may contain 10^{10} nerve cells, whereas the most complex computing machine has only some ten thousand relay units, so how can they possibly be compared? Grey Walter seeks to answer by turning the question round; if direct analysis is virtually impossible, we are forced to adopt an alternative approach. He attempts to set down "without presumption of truth or finality", what may be the lower limit of brain complexity. He suggests that the functional effector units controlled by the brain, multiplied by their degrees of freedom, gives in a crude fashion the number of things the body can be made to do; it seems to be only about 500. The same number roughly applies to the receptor units (the "senses"). However, if we assume all these thousand units are inter-connected with one another, in various permuted systems, the connections would number the order of 10^6, each one of which would produce a distinct action. Grey Walter, in his experiments, has started on the simplest scale, building a moving machine - which he calls his "toy tortoise" - having only 2 effector units (linear and rotary motion) and 2 receptor (by light and touch) and observing that "the behaviour is quite complex and unpredictable". The use of this toy is justified, "not that it does anything particularly well, but that it does anything at all with so little".

One principal method of direct analysis of the workings of the brain, is the electro-encephalographic method; the wave-like rise and fall of potential on the surface of the brain (the "alpha rhythm") first observed by Berger[6] in 1928, have been found to possess a complex structure, which varies with the mental state of the examinee - asleep or awake, relaxed or concentrating etc., particularly with relation to the visual sense[41].† Study of these waveforms, in both normal and epileptic cases, is slowly leading to a detection of "pattern", crudely analogous to the decoding of a cypher by search for its structure, as by counting letter frequencies, etc., although the "cypher" here is overwhelmingly complex. In contrast, the relative simplicity of a computing machine calls for infallibility in its various elements, though as the complexity is increased, some redundancy may perhaps be afforded. In a recent paper[33], Hamming has dealt with the problem of programming a digital machine (which is always liable to make a mistake) using "error detecting" and "error correcting" codes. This facility requires redundancy in the coding, so that the price to be paid is a slightly lower speed of operation.

---

† Dr. David Hartley (1705-1757) suggested that mental phenomena are derived from rhythmic movements in the brain and related the "association of ideas" to these movements.

The possibility of such "self-correction" was first pointed out by Shannon[24] who, as we have already mentioned observes that a coding method may always be found which, by introduction of some redundancy, "is optimal in combatting noise".

One of the most important and humane applications of the "theory of information", which concerns both the biological and the mechanistic fields, is the substitution of one sense for another which has been lost. Early work[4] in this field[¥] includes that of A.M. Bell (Graham Bell's father) who invented a system of "visible speech", for the education of deaf mutes; Braille, invented in 1829, involves the learning of a raised-dot code which employs permutations of 6 positions[ø]. The first machine to convert print directly into sounds was the "Optophone", invented by the Frenchman Fournier d'Albe in 1914, while Naumberg in U.S.A. designed a machine (the "Visagraph") for translating a printed book into embossed characters, unfortunately slow in operation and costly.

The possibility of a machine which can directly read printed type and convert this into intelligible sounds or "feels" is restricted by the fact that the print may be in different sizes and types; this therefore raises the difficult question of Gestalt (perception of form). How do we recognised an object by its shape, irrespective of size and orientation? Or recognise a friend's voice? Or the shape of a ball by its 'feel'? This possibility of sense-replacement is closely dependent upon the amounts of information with which the various senses operate. At first the eye would seem to involve vastly the greatest amount (being, for one thing, "two-dimensional"); however, it would appear that the amount with which the brain has to deal is considerably reduced by the properties of the vision in centring the perceived object, in possessing only a very restricted range of sharp focus and in its accommodation (saturation over areas of uniform colour or brightness, which are "steady-state" and give limited information,[/] in contrast to edges or boundaries). The possibilities of assisting the sense-deficient sufferer are improving and perhaps may soon lead to construction of devices giving greater "rates of information" than those at present in use[37].

(4) SCIENTIFIC METHOD

In this brief history, we have attempted to trace how the idea of information has existed in early times and has gradually entered into a great variety of sciences, to a certain extent integrating them together. Nowadays the concept of information would seem to be essential to all researchers, and as universal and fundamental as the concepts of energy or entropy. Speaking most generally, every time we make any observation, or perform any 'experiment', we are seeking for information; the question thus arises: how much can we know? Or, more specifically, how much can we know from a particular set of observations or experiments? The modern mathematical work, at which we have glanced, seeks to answer in precise terms this very question which, in its origin, is an epistemological one. But firstly, a word of caution. The term "information" has been used by different authors to have different meanings. In previous Chapters we have considered the sense in which it applies to communication engineering ("selective" information) and to analogous fields. The meaning is somewhat different when applied to problems of extraction of "information" from Nature, by experiment.

¥ In U.K. work on the design of reading devices and guiding devices is carried on at St. Dunstans. In U.S.A. work on sense-substitution devices is coordinated under the National Research Council.

ø The earliest raised-character systems for the blind were invented in Spain and France in the 17th cent. (see En. Britt).

/ An analogy is a television signal, passed through a high-pass-filter gives an outline picture, perfectly recognisable.

In a classic work, "The Design of Experiments" (1935) Dr. Fisher[42] considered these problems, largely from the point of view of the application of correct statistical methods[43] and with the subsequent extraction of valid conclusions. The experimenter always assumes "that it is possible to draw valid inferences from the results of an experimentation; that it is possible to argue from consequences to causes, from observation to hypothesis,..... or, as a logician might put it, from the particular to the general". That is, inductive reasoning is involved, essentially, after an experiment has been made: "inductive inference is the only process.....by which new knowledge comes into the world". The experimental method essentially implies uncertainty and the subsequent inductive reasoning raises the thorny question of inverse probability.

In the case of an elementary experiment in, say, mechanics the uncertainty is usually neglected - a body moving in a certain way, under given conditions, will always repeat this motion under the same conditions. But we cannot be so certain in the case of an experiment in, for example, agriculture where the controlling factors are vastly more complex and less well understood. If we are to understand how to draw the maximum information[X] from an experiment then, as Fisher stresses, "the nature and degree of the uncertainty (must) be capable of rigorous expression". The information supplied by an experiment may perhaps be thought of as a ratio of a posteriori to the a priori probabilities.

The importance of the method of inductive reasoning, for arguing from observational facts to theories, seems first to have been recognised (at least in Europe), by the Rev. Thomas Bayes, (1763) who considered the problem -

if $\mathcal{H}_1\ \mathcal{H}_2 \ldots\ \mathcal{H}_i \ldots$ represent various mutually exclusive hypotheses which can explain an event, what are their relative probabilities of being correct? Assume certain data must be known before the event happens, but let E represent some additional data after the event. Then Bayes' Theorem gives

$$P(\mathcal{H}_i / E) = \frac{P(E/\mathcal{H}_i)\, P(\mathcal{H}_i)}{\sum_i P(E/\mathcal{H}_i).\, P(\mathcal{H}_i)}$$

where $P(\mathcal{H}_i / E)$ = probability of $\mathcal{H}_i$ after the event

$P(\mathcal{H}_i)$ = probability of $\mathcal{H}_i$ before the event

$P(E/\mathcal{H}_i)$ = probability of obtaining data E if the hypothesis $\mathcal{H}_i$ be assumed.

X The word "information" is commonly used by statisticians in a special sense. If $P_\theta(x)$ is a distribution function, with some parameter θ (e.g. a mean value) then, writing $L(x/\theta) = \sum_i \log P_\theta(x)$ where $x_1, x_2, x_3 \ldots$ are independent samples, the "information" about θ which these samples give is defined as the mean value of $-\partial^2 L/\partial\theta^2$ However the definition such as is used in communication theory, we have seen to be $-\sum P_\theta(x) \log P_\theta(x)$ where θ is known. This is the "information" given by x and is the mean value of $\log P_\theta(x)$

Although this theorem is generally accepted, its validity is questioned by some mathematicians on the grounds that the prior probabilities $P(\mathcal{H}_i)$ are, strictly speaking, unknown. Bayes' put forward his Axiom, in addition: if there is zero prior data then all hypotheses are to be assumed equally likely, $P(\mathcal{H}_i) = 1/n$. However, an exact knowledge of the probabilities $P(\mathcal{H}_i)$ is relatively unimportant,† as has been stressed by I.J. Good[45] who expresses the above equation logarithmically:-

$$\text{Log } P(\mathcal{H}_i|E) - \text{Log } P(\mathcal{H}_i) = \text{Log } P(E|\mathcal{H}_i) - \text{Log} \sum_i P(E|\mathcal{H}_i) P(\mathcal{H}_i)$$

the extreme right-hand term being a constant. If expected values be taken on both sides (i.e. "averages") this formula may be shown to give Shannon's expression for the rate of transmission of information, R, through a noisy communication channel, in terms of $H(x)$ the (selective) entropy of the input signal and $H_y(x)$ the "conditional entropy" of the input, given the output:-

$$-H_y(x) + H(x) = R \quad \text{or the alternative form:}$$

$$-H_x(y) + H(y) = R$$

Here is then an analogy; an observer receives the distorted output signals (the posterior data E) from which he attempts to reconstruct the input signals (the hypotheses), knowing the language statistics (the prior data).

A recent work by Mackay[46] seeks to obtain a logical quantitative definition of the "information" given by an experiment or scientific proposition. He observes: "many scientific concepts in different fields have a logically equivalent structure. One can abstract from them a logical form which is quite general, and takes on different particular meanings according to the context..... It is suggested that the most fundamental abstract scientific concept is "quantal" in its communicable aspects". Mackay makes reference to the work of Wittgenstein[47] who, in his "Tractatus Logico Philosophicus" (1922) considers the conditions which must be fulfilled by a logically perfect language in order to described "the world" (i.e. "experience"). Such descriptions are thought of as consisting of complexes of facts; the simplest kind of fact is called an "atomic fact" and can be answered by true or false. Such facts cannot be further reduced to component atomic facts. Of course such statements only have meaning in an agreed "universe of discourse" a point which per - haps Mackay does not sufficiently emphasise. This formal logical view is applied to scientific concepts, which are based on limited data given by sets of observations, and with the conclusion that a scientific state- ment may be dissected into elementary ("atomic") propositions, each of which may be answered by true or false; a "unit of information" is then defined as that which decides us to add one elementary proposition to the logical pattern of the scientific statement. Mackay then draws attention to two complementary aspects of information. Firstly, the a priori aspect, related to the structure of the experiment - for example a galvanometer may perhaps have a response time of 0.01 secs; to describe readings at closer intervals is impossible, the instrument is capable only

---

† See comment by Dr. Woodward, page

of giving information in terms of these small, but finite (quantal) intervals. This structural aspect corresponds to the logon concept of Gabor[15], originally framed to define the response characteristics of a communication channel (see Sec(2)). Experimentation abounds with similar uncertainties: "each time that a compromise has to be struck, say between the sensitivity and response time of a galvanometer, or the noise-level and band-width of an amplifier, or the resolving power and aperture of a microscope....." Secondly the a posteriori aspect, related to the "metrical information - content" of the experiment; for example the galvanometer may be used to record a set of values of a magnitude each reading representing a certain "amount of metrical information". These recordings being capable of only a certain accuracy, the amount of metrical information obtained may be thought of as a dimensionless measure of precision, or weight of evidence. It is related to Fisher's definition[42] of amount of statistical information, as the reciprocal of the variance of a statistical sample.

Mackay represents the information content of a result by means of a vector in multi-dimensional "information-space"; the number of dimensions and the square of the vector length indicating respectively the amounts of these a-priori and a posteriori features of information provided. The various uncertainty-relations of physics "appear basically as axioms expressing the quantal nature of communicable information, consequent on the use of logical forms; and the quantity entropy plus information-content appears as a fundamental invariant of a physical system".

Referring to his book "Cybernetics", Wiener observes: "One of the lessons of the book is that any organism is held together by the possession of means for the aquisition, use, retention and transmission of information"; in this short history we have collected some evidence. Such means exist in human society in terms of their spoken language, their press, telephone system, and so on, enabling it to operate as an integrated organism; and insect society is linked by its language of smells, postures or sounds. The human body can operate as a unit in as much as it possesses means of communicating information, in the brain and nervous system. Most generally, any physical experiment or proposition is an entity, in that it involves communication of information between the observer and the observed and in that this can be given a logical structure.

In conclusion, the author would like to acknowledge, with gratitude, the assistance given by numerous friends, especially Dr. D. Gabor, in the compiling of this short history.

CONCLUDING REMARKS:-

The remarks made by various speakers who attended the Symposium, and certain others in private correspondence, have been considered in the preparation of the present version of this paper; since this is intended as a history it has been felt that this procedure would render it a more accurate chronicle.

REFERENCES

Languages and Codes

1. Diringer D., "The Alphabet" Hutchinson's Ltd., London, 1948.

2. Bodmer F., "The Loom of Language" (Edited by L. Hogben) Allen & Unwin, London, 1944.

3. Pratt F., "Secret and Urgent", Blue Ribbon Books (1939), (for letter frequency tables).

4. Encyclopaedia Britannica.

5. D'Agapayeff "Codes and Ciphers" O.U.P. 1939.

6. "Science Since 1500" (H.M. Stationery Office) London, 1938. See also Refs. 24, 29, 33.

Communication Theory

7. Colpitts E.H. and Blackwell O.B., "Carrier Current Telephony and Telegraphy", A.I.E.E. Trans. Vol. 40, 1921, p.205.

8. Pero Meade S., "Phase Distortion and Phase Distortion Correction", B.S.T.J. Vol.7, April 1928, p.195.

9. Carson J., "Notes on the Theory of Modulation" Proc. I.R.E., Vol.10, Feb. 1922, p.57.

10. Nyquist H., "Certain Factors Affecting Telegraph Speed", B.S.T.J., Vol.3, April 1924, p.324.

11. Küpfmüller K., "Transient Phenomena in Wave Filters", Elek. Nach. Tech., Vol.1, 1924, p.141.

12. Hartley R.V.L., "Transmission of Information", B.S.T.J., Vol.7, Aug.1928, p.535.

13. Balth van der Pol, "Frequency Modulation", Proc. I.R.E., Vol.18, No.7, July 1930, p.1194.

14. Armstrong E.H., "A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation", Proc. I.R.E. Vol.24, May 1936, p.689.

15. Gabor D., "Theory of Communication", Jour. I.E.E. Vol.93, Part III, Nov. 1946.

16. Ville, J.A., "Théorie et Applications de la Notion de Signal Analytique", Cables et Transmission, 1948, p.61.

17. Paget, Sir Richard, "Human Speech", Kegan Paul, Trench, Trubner and Co. Ltd., 1930.

18. Fletcher Harvey, "Speech and Hearing", Van Nostrand, 1929.

19. Dudley Homer, Riesz R.R., Watkins S., "A Synthetic Speaker", Jour. Franklin Inst. 1939, p.739.

20. Dudley Homer, "Remaking Speech", Jour. Acoust. Soc. Amer. 1939, Vol.11, p.169.

21. Halsey R.J. and Swaffield J., "Analysis-Synthesis Telephony, with Special Reference to the Vocoder", Jour. I.E.E., Part III, Sept. 1948.

22. Gabor D., "New Possibilities in Speech Transmission", Jour. I.E.E., Part III, Nov. 1947 and Vol.95, Part III, Sept. 1948.

23. Tuller W.G., "Theoretical Limits on the Rate of Transmission of Information", Proc. I.R.E., Vol.37, May 1949, p.468.

24. Shannon C.E., Bell Syst. Tech. Jour. Vol.27, pp. 379 and 623.
Shannon C.E. and Weaver W., "A Mathematical Theory of Communication", University of Illinois Press, Urbana, 1949.

25. Cooke D., Jelonek Z., Oxford A., Fitch E., "Pulse Communication", Jour. I.E.E. IIIa, 1947, p.83.

26. Deloraine E.M., "Pulse Modulation", Proc. I.R.E., June 1949, p.702.

27. Fano R.M., "The Transmission of Information", M.I.T. Tech. Rep. No.65, March 17, 1949.
See also Refs. 3, 37, 48.

Automatic Digital Computing Machines

28. "Proceedings of a Symposium on Large-Scale Digital Calculating Machinery", The Annals of the Computation Laboratory of Harvard University, Vol.XVI (O.U.P. in the U.K.)

29. "Report of a Conference on High Speed Automatic Calculating-Machines" University Mathematical Laboratory, Cambridge, England, Jan. 1950. (Ministry of Supply).
(See this for extensive bibliography).

30. Hartree D.R., "Calculating Instruments and Machines", University of Illinois Press, Urbana, 1949.

31. Hartree D.R., "An Historical Survey of Digital Calculating Machines", Proc. Roy. Soc., Vol.195, 1948, p.265.

32. Babbage H.P., "Babbage's Calculating Engines", Spon Ltd., London, 1889.
See also Ref. 33, 34, 36, 37.

Aspects of "Mechanised Thinking"

33. Hamming R.W., "Error Detecting and Error Correcting Codes", B.S.T.J., Vol.29, April 1950, p.147.

34. McCulloch W.S. and Pitts W., "A Logical Calculus of the Ideas Immanent in Nervous Activity", Bull. Math. Biophys. 5, 1943, p.115.

35. J. von Neumann and Morgenstern "Theory of Games", Princeton, 1947.

36. Shannon C.E., "Programming a Computor for Playing Chess", Phil. Mag., Vol.41, March 1950, p.256.

See also (a) Shannon C.E., "Communication Theory of Secrecy Systems" B.S.T.J., Vol.28, Oct. 1949, p.656.
(b) Edmund Berkeley, "Giant Brains, or Machines that Think".
(c) Refs. 6, 29, 37, 41.

Feedback, Controlled Systems, "Cybernetics"

37. Wiener N., "Cybernetics", John Wiley, 1948.

38. Black H.S., "Stabilised Feedback Amplifiers", Elec. Eng. Vol.53, 1934, p.114.

39. Nyquist H., "Regeneration Theory", B.S.T.J., Jan. 1932, p.126.

40. Bode H.W., "Network Analysis and Feedback Amplifier Design", Van Nostrand Co., 1945.

41. Walter Grey, "The Functions of Electrical Rhythms in the Brain" (24th Maudsley Lecture), Jour. of Mental Science, No.402, Vol.XCVI, Jan. 1950.

See also Ref. 4, under "Behaviourism".

Information Theory and Scientific Method

42. Fisher R.A., "The Design of Experiments", Oliver and Boyd, London, 1935.

43. Fisher R.A., "Statistical Methods for Research Workers", Oliver and Boyd, London, 1925.

44. Bayes T., "An Essay towards Solving a Problem in the Doctrine of Chances", Phil. Trans. Roy. Soc., I, iii, 370 (1763).

45. Good I.J., "Probability and the Weighing of Evidence", Griffen and Co., 1950.

46. Mackay D.M., "Quantal Aspects of Scientific Information", Phil. Mag. Vol. 41, March 1950, p.289.

47. Wittgenstein L., "Tractatus Logico-Philosophicus", 1922, Kegan Paul.

48. Szilard L., "Uber die Entropieverminderung in einem Thermodynamischen System bei Eingriffen Intelligenter Wesen". Zeit. f. Phys., Vol.35, 1929, p.840.

Noise

49. Einstein A., "Theory of Brownian Movement", Methuen, 1926.

50. Schottky W., "Über Spontane Stromswarkungen in verschiedenen Elektrizitätsleiten" Ann. d. Phys., Vol.57, 1918, p.541.

51. Schottky W., "Zur Berechnung und Beurteilung des Schotteffektes", Ann. d. Phys. Vol.68, 1922, p.157.

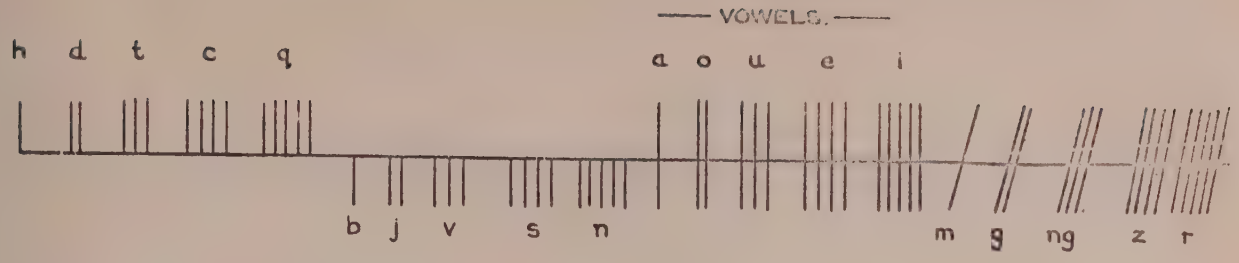
52. Johnson J.B., "Thermal Agitation of Electricity in Conductors", Phys. Rev., Vol.32, 1928, p.97.

53. Nyquist H., "Thermal Agitation of Electric Charge in Conductors", Phys. Rev., Vol.32, 1928, p.110.

54. Macdonald D.K.C., "Some Statistical Properties of Random Noise" Proc. Camb. Phil. Soc., Vol.45, 1949, p.368.

55. S.O. Rice, "Mathematical Analysis of Random Noise", B.S.T.J., Vol.23 (1944) p.282 and Vol.24 (1945) p.46.

FIG. 3.



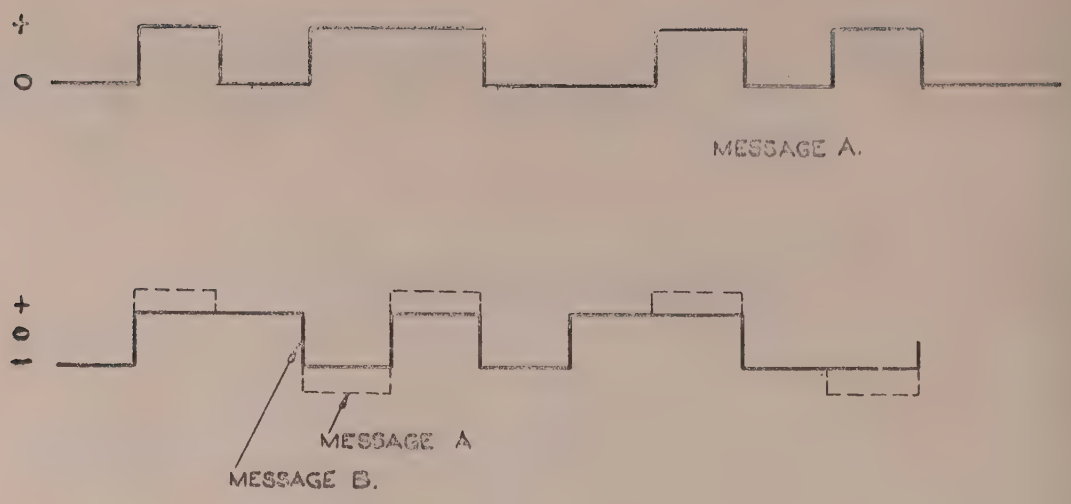
THE OGAM (CELTIC) SCRIPT.

FIG. 4.

E	‒	131·05	PER 1000.
T	———	104·68	
A	‒ ———	81·51	
I	‒ ‒	63·45	
N	——— ‒	70·98	
O	——— ——— ———	79·95	
S	‒ ‒ ‒	61·01	
H	‒ ‒ ‒ ‒	52·59	
R	‒ ——— ‒	68·32	
D	——— ‒ ‒	37·88	
L	‒ ——— ‒ ‒	33·89	
U	‒ ‒ ———	24·59	
C	——— ‒ ——— ‒	27·58	
M	——— ———	25·36	
F	‒ ‒ ——— ‒	29·24	
W	‒ ——— ———	15·39	
Y	——— ‒ ——— ‒	19·82	
G	——— ——— ‒	19·94	
P	‒ ——— ———	19·82	
B	——— ‒ ‒ ‒	14·40	
V	‒ ‒ ‒ ———	9·19	
K	——— ‒ ———	4·20	
Q	——— ——— ‒ ———	1·21	
J	‒ ——— ——— ———	1·32	
X	——— ‒ ‒ ———	1·66	
Z	——— ——— ‒ ‒	0·77	

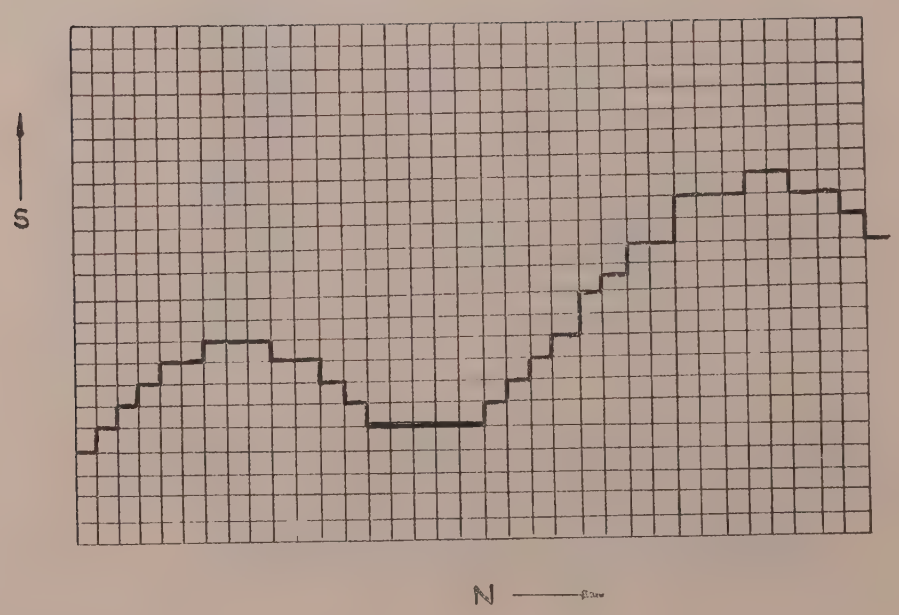
MORSE CODE AND LETTER FREQUENCIES.

SYMBOLS IN ORDER OF PROBABILITY AS IN MORSE'S DAY; THE LETTER FREQUENCIES ARE THOSE OF MODERN ENGLISH.



EDISON'S SYSTEM OF DUPLEX.

FIG. 6.



"QUANTISING" OF A CONTINUOUS WAVE.

COMMUNICATION THEORY - EXPOSITION OF FUNDAMENTALS

by
C.E. Shannon

ABSTRACT

In any branch of applied mathematics, the vague and ambiguous concepts of a physical problem are given a more refined and idealized meaning. In information theory, one of the basic notions is that of the amount of information associated with a given situation. "Information" here, although related to the everyday meaning of the word, should not be confused with it. In everyday usage, information usually implies something about the semantic content of a message. For the purposes of communication theory, the "meaning" of a message is generally irrelevant; what is significant is the difficulty in transmitting the message from one point to another.

From this point of view, information exists only when there is a choice of possible messages. If there were only one possible message there would be no information; no transmission system would be required in such a case, for this message could be on a record at the receiving point. Information is closely associated with uncertainty. The information I obtain when you say something to me corresponds to the amount of uncertainty I had, previous to your speaking, of what you were going to say. If I was certain of what you were going to say, I obtained no information by your saying it.

In general, when there are a number of possible events or messages that may occur, there will also be a set of a priori probabilities for these messages and the amount of information, still arguing heuristically, should depend upon these probabilities. If one particular message is overwhelmingly probable, the amount of information or the a priori uncertainty will be small.

It turns out that the appropriate measure for the amount of information when a choice is made from a set of possibilities with the probabilities $p_1, p_2, \cdots, p_n$ is given by the formula

$$H = - \sum_{i=1}^{n} p_i \log p_i \, . \qquad (1)$$

Some of the reasons justifying this formula are (1) $H = 0$ if and only if all the p_i are zero except one which is unity, i.e., a situation with no choice, no information, no uncertainty. (2) With a fixed n, the maximum H occurs when all the p_i are equal, $p_i = \frac{1}{n}$. This is also, intuitively, the most uncertain situation. H then reduces to log n. (3) H is always positive or zero. (4) If there are two events x and y, we can consider the information H_c in the composite event consisting of a choice of x and y.

$$H_c(x, y) = - \Sigma \, p(x, y) \log p(x, y) \, . \qquad (2)$$

It can be shown that this composite information is greatest when the two events, x and y, are statistically independent. It is then the sum of the individual amounts of information.

Equation (1) is identical in form with certain formulas for entropy used in statistical mechanics, in particular in the formulation due to Boltzmann. It is to be noted that both here and in thermodynamics $- \Sigma p_i \log p_i$ is a measure of randomness, in thermodynamics, the random position of a representative point in a dynamical phase-space, in information theory the randomness in the choice of the particular message to be transmitted from an ensemble of possible messages. We shall frequently speak of quantities having the form $- \Sigma p_i \log p_i$ as entropies because of this identity in form.

The formula (1) measures the amount of information when a single choice is made from a finite set of possible events. In a communication system we frequently must consider messages which are produced by a sequence of such choices. Thus, the English text to be transmitted over a telegraph system consists of a sequence of letters, spaces and punctuation. In such a case, we are concerned with the amount of information produced per symbol of text. The formula (1) must be generalized to take account of influences between letters and the general statistical structure of the language. We think of a language, then, as being produced by a stochastic (i.e., statistical) process which chooses the letters of a text one by one in accordance with certain probabilities depending in general on previous choices that have been made.

Samples of statistical English based on such a representation of the English language have been constructed. The following are some examples with varying amounts of the statistics of English introduced.

1. Letter approximation (letter probabilities the same as in English)

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI

2. Trigram approximation (probabilities for triplets of letters the same as in English)

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID

3. Word-digram approximation (probabilities for word-pairs as in English)

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER
   THAT THE CHARACTER OF THIS POINT IS THEREFORE

4. Word-tetragram approximation

   THIS WAS THE FIRST. THE SECOND TIME IT HAPPENED
   WITHOUT HIS APPROVAL. NEVERTHELESS IT CANNOT BE
   DONE. IT COULD HARDLY HAVE BEEN THE ONLY LIVING
   VETERAN OF THE FOREIGN POWER HAD STATED THAT NEVER
   MORE COULD HAPPEN.

The amount of information produced by a stochastic process per letter of message is defined by formulas similar to (1). For example, one method is to calculate the amount of information for a choice of N letters of text, divide by N to put it on a per letter basis, and then allow N to increase indefinitely.

The fundamental reason why the entropy per letter obtained in this way forms the appropriate measure of the amount of information is contained in what may be called the "coding theorem." This states that if a language has an entropy H bits per letter (i.e., $\log_2$ was used in the calculation) then it is possible to approximate as closely as desired to a coding system which translates the original messages into binary digits (0 or 1) in a reversible way and uses, on the average, H binary digits in the encoded version per letter of the original language. Furthermore there is no such system of encoding which uses less than H binary digits on the average. In other words, speaking roughly, H measures the equivalent number of binary digits for each letter produced in the language in question. H measures all languages by the common yardstick of binary digits.

A closely related aspect of a language is its redundancy. This is defined as follows. Suppose all the letters in the language were independent and equiprobable. Then the entropy per letter would be the logarithm of the number of letters in the alphabet. The relative entropy is the ratio of the actual entropy to this maximum possible entropy for the same alphabet. The redundancy is one minus the relative entropy. The redundancy determines how much a language can be compressed when properly encoded into the same alphabet. Thus, if the redundancy were 70 per cent, a suitable encoding of the language would reduce its length on the average by this amount.

A number of methods have been developed for estimating the entropy and redundancy of various stochastic processes. In the case of printed English, the most direct approach is to make use of tables of letter, digram, trigram, etc., probabilities and to calculate from them the entropy of the various approximations to English. Unfortunately, with the tables actually available it is not possible to go farther than approximations including about six or eight letters. At this point, figures of the order of 50 per cent for redundancy are obtained. They of course do not include long range statistical influences extending over groups of words, phrases and sentences.

Another more delicate method of estimating these parameters has recently been devised. It is based on the fact that anyone speaking a language possesses implicitly an enormous knowledge of the statistical structure of that language. By a relatively simple experiment, it is possible to translate this knowledge into numerical data which give upper and lower bounds for the entropy and redundancy. The experiment is to ask a subject to guess an unknown text in the language letter by letter. At each letter he guesses first what he considers the most probable next letter in view of the preceding text. If he is wrong he is required to guess again, and so on until he finally arrives at the correct next letter. In a typical experiment of this type with a text containing 102 letters, the subject guessed right on his first guess 79 times. Eight times he was right on the second guess, three times on the third, twice each on the fourth and fifth, and only eight times required more than five guesses. These figures clearly indicate the great redundancy of English. Furthermore from them one can estimate upper and lower numerical bounds for the redundancy which take into account rather long-range structure, inasmuch as the subject made considerable use of this structure in formulating his guesses. From the results of this work it appears that the redundancy of printed English at 100 letters is of the order of 75 per cent, and may well exceed this figure for still longer range structure.

So far we have been considering information only in the discrete cases. In generalizing to the continuous case, for example a speech wave or a television signal, a number of new features emerge. The generalization is by no means trivial. In the first place, a continuously variable quantity is capable of assuming an infinite number of possible values, and if there were no other considerations this would imply an infinite amount of information. Actually in practical cases there are always features which prevent this and enable one to effectively reduce the continuous case to a discrete case. The two facts which produce this result are the presence of perturbing noise in the signal and the finite resolving power of any physical receiving apparatus.

One important mathematical result which expedites the analysis of continuous information is the "sampling theorem." This states that a function of time limited in frequency components to a band W cycles wide is determined by giving its values at a series of sample points equally spaced in time and separated by $\frac{1}{2W}$ seconds. The knowledge of such a function is equivalent to knowledge of a sequence of numbers, the numbers occurring at the rate of 2W per second. If a message consists of such a band-limited function of time which persists for substantially T seconds, it is determined by giving 2TW numbers. Geometrically, such a function can be represented by a point in a space with 2TW dimensions. Certain aspects of communication theory can be analysed by a consideration of the properties of mappings (which correspond to systems of modulation) in such spaces.

The problems of measuring the amount of information in a continuous message is more involved than a simple generalization of the entropy formula. It is necessary at this point to introduce a measure of the fidelity of reception of the message when it is perturbed by noise. When a suitable measure of fidelity has been set up, it is possible to define the amount of information (in bits per second) for a given continuous source and for a given fidelity of transmission. As the fidelity requirements are made more stringent, the amount of information increases. For example, in transmitting

English speech, if we are satisfied with an intelligible reproduction the amount of information per second is small; if a high fidelity reproduction is required, preserving personal accents, etc., the information is greater.

REFERENCES:

1. Shannon, C.E. and Weaver, W. "The Mathematical Theory of Communication", University of Illinois Press, Urbana, 1949.

2. Shannon, C.E. "Communication in the Presence of Noise", Proc.Inst. of Radio Engineers, 37, pp.10-21, January 1949.

3. Shannon, C.E. "The Prediction and Entropy of Printed English", to appear in Bell System Technical Journal, January 1951.

COMMUNICATION THEORY AND PHYSICS.

by D. Gabor

SUMMARY

The electromagnetic signals used in communication are subject to the general laws of radiation. One obtains a complete representation of a signal by dividing the time-frequency plane into cells of unit area and associating with every cell a "ladder" of distinguishable steps in signal intensity. The steps are determined by Einstein's law of energy fluctuation, involving both waves and photons.

This representation, however, gives only one datum per cell, viz. the energy, while in the classical description one has two data; an amplitude and a phase. It is shown in the second part of the paper that both descriptions are practically equivalent in the long-wave region, or for strong signals, as they contain approximately the same number of independent, distinguishable data, but the classical description is always a little less complete than the quantum description. In the best possible experimental analysis the number of distinguishable steps in the measurement of amplitude and phase is only the fourth root of the number of photons. Thus it takes a hundred million photons per cell in order to define amplitude and phase to one percent each.

Communication theory has up to now developed mainly on mathematical lines, taking for granted the physical significance of the quantities which figure in its formalism. But communication is the transmission of physical effects from one system to another, hence communication theory should be considered as a branch of physics. Thus it is necessary to embody in its foundations such fundamental physical data as the quantum of action, and the discreteness of electric charges. This is not only of theoretical interest. With the progress of electrical communications towards higher and higher frequencies we are approaching a region in which quantum effects become all-important. Nor must one forget that vision, one of the most important paths of communication, is based essentially on quantum effects.

Some years ago I have proposed a mathematical framework for the representation of signals.(1) I have been rightly criticized for having left out noise, which is an essential feature of all communications. This will be remedied here, and at the same time the description will be brought in line with modern physics. But as the mathematical frame will serve as a useful foundation, it will be necessary to give first a short review of it.

(1) CLASSICAL REPRESENTATION OF SIGNALS

The previous work(1) started from the observation that the description of a signal in the conventional way, as a continuous function of time is redundant and non-physical. A continuous function contains in any interval, however small, an infinity of data, corresponding to an infinite range of frequencies. A similar objection can be raised against the Fourier representation, which involves infinite time. In the new description one considers the signal simultaneously as a function of frequency and of time. It is convenient to use only positive frequencies in the description. This can be done by introducing a certain complex function, whose real part is the physical signal. (The theory of these complex or "analytical" signals has in the meantime received interesting additions by the work of J.A. Ville.(2)) If the time-frequency halfplane, Fig. 1, is divided, by any network, into cells of unit area, $\Delta t \, \Delta \nu = 1$, one finds that the signal in any domain containing a sufficiently large number of cells is fully described by associating two real data, or one complex datum with every cell. In other words, each cell has two degrees of freedom.

One can represent an arbitrary signal in an infinity of ways as a linear function of certain "elementary" signals, associated with the individual cells. There is however one description of particular interest, in which the elementary signals are harmonic functions, modulated with a "gaussian" signal, i.e. they have envelopes of probability shape. (Fig. 2). These share with other functions the property that their Fourier transforms have the same shape, but they are unique in that respect that the product of their "effective" duration and of their effective frequency width is the smallest possible for any function. Thus it can be said that these elementary functions overlap as little as possible. They have also the advantage that the familiar concepts "amplitude" and "phase" can be used in connexion with them in the same way as with infinite harmonic functions. The complex elementary function is COS + j SIN, if we denote by COS the even, and by SIN the odd type of real elementary signal. Multiplying these with suitable complex coefficients c_{ik}, as indicated in Fig. 1, we can represent any arbitrary signal. Time-description and frequency-description (Fourier integral,) can be considered as extreme special cases of this representation. In the first case the even elementary signal degenerates to a delta-function, and the odd one to its derivative, in the second case both become ordinary harmonic functions.

The "matrix" representation, illustrated in Fig. 1, is proof against the objections raised against the pure "time" or "frequency" descriptions, but it does not go far enough. The infinity of data has been reduced to a finite number, but these data, i.e. the coefficients c_{ik} are still supposed to be exactly defined. But a single exact datum still contains an infinite amount of information, i.e. an infinite number of "yesses or noes". In reality of course these amplitudes, like every physical datum, have a certain amount of uncertainty or "noise". This has been taken into consideration in the mathematical theories of Shannon 3 and Tuller 4, where the noise amplitudes, or certain functions of it are assumed as known. But we cannot be satisfied with this in a physical theory. Even if all accidental imperfections of the instruments are eliminated, there remain certain basic uncertainties, which we are now going to investigate.

(2) STATISTICAL PROPERTIES OF THE INFORMATION CELL.

In order to connect the mathematical scheme with physical reality we must first observe, that every physical signal has a certain energy associated with it. We can also associate a certain energy with every cell*, and for simplicity we will say that it is "contained" in the cell.

We will limit the discussion to electric communications, though most of our results will apply equally well to sound, or, in short, to communication by any quantity which is considered as continuous in classical physics. We observe first that all electric signals are conveyed by radiation. Even if lines or cables are used in the transmission, by the Maxwell-Poynting theory the energy can be located in empty space. Hence we can apply to our problem the well known results of the theory of radiation.

For simplicity we consider our communication system as having the uniform temperature T. The uncertainty connected with the concept of temperature will produce certain fluctuations in the energy of the cells, which we can calculate by the rules of statistical thermodynamics once we know the law for the mean thermal energy $\bar{\varepsilon}_T$ of a cell, in function

---

*The energies of the elementary signals of the type discussed will not add up exactly to the energy of the whole signal, because they are not quite orthogonal, but the error will vanish in the mean over large numbers of cells. One can however, to meet objections, consider instead an orthogonal set of elementary signals, such as the signals with "limited spectrum", introduced by Shannon (5) and by Oswald (6) which have uniform spectral density inside and zero outside a frequency strip. This makes no difference to the following discussion, as no reference will be made to any special type of elementary signal.

of the temperature. This we obtain at once, if we observe that every cell in the signal has two degrees of freedom, of which only one counts for the purpose of statistics, as the other is of the nature of a phase. Thus, by Planck's law

$$\bar{\varepsilon}_T = \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1} \qquad (1)$$

where ν is the mean frequency of the cell. In other words we identify every information cell with a "Planck oscillator".

This requires perhaps a little more explanation, as physicists are less familiar with a discussion of radiation in terms of frequency and time than in terms of frequency and space. But the first case is immediately reduced to the second if we imagine the signal propagating with the velocity c, and plot the information diagram against the length ct instead of against the time t. The state of the field in such a linear system, (in which we consider one state of polarization only,) can be represented by two systems of progressive waves in opposite directions, only one of which represents the signal in which we are interested. We count the degrees of freedom in this wave system, - which is what we have done in the last section,- and give it the energy $\bar{\varepsilon}_T$ for each free amplitude, disregarding the phases. This is the application to the linear case of v. Laue's well known derivation of Planck's law for the radiation density by superposition of plane waves.(7)

The thermal energy $\bar{\varepsilon}_T$ does not in itself represent "noise" i.e. an uncertain disturbing factor. It becomes disturbing only by its fluctuations. In order to obtain the mean square fluctuations of the energy, we apply Einstein's formula

$$\overline{\delta \varepsilon_T^2} = \overline{\left(\varepsilon_T - \bar{\varepsilon}_T\right)^2} = kT^2 \frac{d\bar{\varepsilon}_T}{dT} \qquad (2)$$

which gives

$$\overline{\delta \varepsilon_T^2} = h\nu\bar{\varepsilon}_T + \bar{\varepsilon}_T^2 \qquad (3)$$

Einstein's interpretation of this equation is well known.(8) The second term has been identified by H.A. Lorentz with the fluctuations due to the interference of waves with random phases*. But the first term suggests that the energy is concentrated in light quanta or photons of energy $h\nu$ which fluctuate in any element as if

---

*In technical theories of thermal noise it is usually forgotten that it is not the noise power but its fluctuations which cause the disturbance. But if the quantum effect is small and the second term in eq. 3 predominates, the r.m.s. value of the noise power fluctuation is equal to the noise power itself, hence this error is without consequences.

From eq. 3, neglecting the quantum term one can easily derive Nyquist's well known rule (9), that a resistance R can be considered as containing a "noise generator" with a mean square electromotive force

$$\overline{E^2} = 4kTR\,\Delta\nu$$

The proof can be given in the same form as Nyquist has done, by substituting a cable with wave impedance R for the resistance. But one must add the condition that not only the noise power, but also its fluctuations follow the same rule in the resistance as in the cable. This is no arbitrary rule, the necessity of a uniform law of mean square fluctuations follows directly from the second principle, as Szilárd (10) has shown.

they were independent particles. As both effects are present simultaneously, and are acting as if they were independent of one another, it is not possible to make a simple physical picture of the process. But fortunately this is not necessary. Applying Einstein's argument to the case when a signal is present, so that the mean energy $\bar{\varepsilon}$ in the cell exceeds the thermal mean energy $\bar{\varepsilon}_T$ we obtain in Appendix I.

$$\overline{\delta\varepsilon^2} = \overline{(\varepsilon - \bar{\varepsilon})^2} = h\nu\,\bar{\varepsilon} = 2\,\bar{\varepsilon}\,\bar{\varepsilon}_T - \bar{\varepsilon}_T^{\,2} \tag{4}$$

Expressing the energy by the number of photons N in the cell, so that $\varepsilon = N\,h\nu$, $\bar{\varepsilon}_T = \bar{N}_T\,h\nu$, this becomes

$$\overline{\delta N^2} = \overline{(N - \bar{N})^2} = \bar{N}(1 + 2\bar{N}_T) - \bar{N}_T^2 \tag{5}$$

with

$$\bar{N}_T = 1/(e^{\frac{h\nu}{kT}} - 1) \tag{6}$$

$\bar{N}_T$, the number of "thermal" photons per cell, is a large number for the frequencies used in electrical communications. One can use the approximate formula

$$\bar{N}_T = \frac{kT}{h\nu} = \frac{kT}{hc}\lambda = 0.7\,\lambda T$$

which gives $\bar{N}_T = 210$ for a wavelength of 1 cm and a temperature of 300° K. On the other hand for visible light, $\lambda = 5.10^{-5}$ cm, the approximate formula

$$\bar{N}_T = e^{-\frac{h\nu}{kT}} = e^{-1/0.7\lambda T}$$

gives $\bar{N}_T = e^{-94} = 10^{-39}$. Hence for visible light, at ordinary temperatures there is practically no thermal noise, and the fluctuation becomes pure "quantum noise", which follows the law $\overline{\delta N^2} = \bar{N}$.

These results enable us to construct a complete physical representation of a signal. We see that the state of an information cell is completely determined by the stochastic number N, the number of photons. We can now construct a scale or "ladder" of distinguishable states, on which every step corresponds to a reasonable ascertainable difference. It is an evident suggestion to adopt the r.m.s. fluctuation of N as the unit step.* With this convention, the number of steps distinguishable below a maximum level N_m is, by eq. 3

$$S = \int_{\bar{N}_T}^{N_m} \frac{dN}{(\overline{\delta N^2})^{\frac{1}{2}}} = \int_{\bar{N}_T}^{N_m} \frac{dN}{\left[N\,(1 + 2\bar{N}_T) - \bar{N}_T^2\right]^{\frac{1}{2}}} =$$

$$= \frac{2}{1 + 2\bar{N}_T}\left\{\left[N_m\,(1 + 2\bar{N}_T) - \bar{N}_T^2\right]^{\frac{1}{2}} - \left[\bar{N}_T\,(1 + \bar{N}_T)\right]^{\frac{1}{2}}\right\} \sim \frac{2N_m^{\frac{1}{2}}}{(1 + 2\bar{N}_T)^{\frac{1}{2}}} \tag{7}$$

---

* By the theorem of Bienaymé and Tchebycheff the probability of an error k times exceeding the r.m.s. error or fluctuation is smaller than $1/k^2$, whatever the law of the fluctuations may be.

The last formula is valid for large signals.

In the useful terminology introduced by D.M. MacKay(11) S is the "proper scale" of the photons. Fig. 3 illustrates the representation of a signal in three dimensions, with the photon scale at right angles to the time-frequency plane. If instead we plotted $\log_2 S$, the ordinates would give directly the equivalent number of binary selections or "bits". This is the number of "yesses or noes" required to fix the position of the signal on the ladder.

It is of interest to enquire about the minimum energy required for the transmission of the first "bit" of information. By our convention this cannot be less than the r.m.s. value of the thermal energy fluctuations, though it can be more. Combining eqs. 1 and 3 one obtains

$$\varepsilon_{min} = (\overline{\delta\varepsilon_T^2})^{\frac{1}{2}} = \frac{h\nu}{2 \sinh \frac{h\nu}{2kT}} \qquad (8)$$

But on the other hand this energy cannot be less than one quantum, $h\nu$. Thus the energy required for the first "bit" is either ε_{min} as given by eq. 8, or $h\nu$, whichever is the larger of the two. As shown in Fig. 4, the two lines cross at $h\nu_0 = 0.96$ kT, which in only slightly less than kT. Thus up to a critical frequency ν_{cr} an energy kT is sufficient for the first step, but no communication is possible with an energy of less than kT. The interesting feature of this result is its generality, it applies to the unknown processes in the nervous system as well as to electrical communications.

(3) STATISTICS OF SIGNALS.

Up to the point we have directed our attention only to one cell. A short discussion of signals and of ensembles of signals may not be out of place before we return to the physical analysis of our results.

A signal is a system of, say, n cells, preferably but not necessarily contiguous in the information plane. We consider a large number of such systems in a stationary transmission, in which they differ only by their position in time, not in frequency, and we speak of this number as of an "ensemble". Evidently the analogy with statistical mechanics is not very perfect. In statistical mechanics we can either follow a system of an ensemble over a long time, or we can look at all systems in the ensemble simultaneously, while here we can take only the first view. It is also somewhat questionable whether the often used expression "ergodic" is justified. In its original sense, due to Willard Gibbs, it means that each system, /each signal,/ spends equal times in all states compatible with a given energy, which is not true for most stationary transmissions usually considered as "ergodic" in communication theory. Thus we prefer to avoid this term.

The most important mean value in such an ensemble is that of the entropy. In order to clarify the connexions between physics and communication theory, it may be useful to consider this problem in two stages. In the first we consider all configurations of the system of n cells, compatible with the energetic conditions of the transmission as equally probable, and define the entropy k log P, where P is the number of all these possible configurations. In the second stage, however, we give them different probabilities or "weights". The first stage is in close connexion with physics, actually it is the problem of calculating the entropy of an "ergodic" system, in the original meaning of the word. According to quantum statistics all simple, accessible states have equal probability, and the levels of Planck oscillators are of course simple. (Cf. Jordan(12)).

As an example let us estimate the mean entropy of a system of n cells in a transmission in which the mean energy level is $\overline{S^2}$, and the r.m.s. deviation from the mean is ΔS_n^2, for brevity. (The suffix n had

to be added, as the standard deviation is dependent on the size of the sample.) For simplicity we assume $\Delta S_n^2 \ll \overline{S^2}$, i.e. a small degree of "modulation". This allows us to equate the number P of possible states to the number of points with positive, integer coordinates inside a sherical shell in n-dimensions, whose mean radius R is given by

$$R^2 = n\,\overline{S^2}$$

while its thickness is

$$2\Delta R = n\Delta S_n^2/R$$

Well known formulas for the volume of a n-dimensional sphere give

$$P = \frac{(\pi/4)^{\frac{1}{2}n}}{\Gamma\left(\frac{1}{2}n+1\right)}\, nR^{n-1}.2\Delta R = \frac{(\pi/4)^{\frac{1}{2}n}}{\Gamma\left(\frac{1}{2}n+1\right)}\, n(n\overline{S^2})^{\frac{1}{2}n}\,\frac{\Delta S_n^2}{\overline{S^2}} \approx \frac{1}{\sqrt{\pi}}\left(\tfrac{1}{2}e\pi\overline{S^2}\right)^{\frac{1}{2}n}\frac{\sqrt{n}\Delta S_n^2}{\overline{S^2}} \qquad (9)$$

We have used Stirling's formula to approximate the Γ - function for large arguments. For sufficiently large n, however complicated the signals, they must behave as if they were independent, and we must have asymptotically

$$n.\ (\Delta S_n^2)^2 \longrightarrow \text{const.}$$

hence, apart from a constant

$$S \longrightarrow k\ n.\log\ (\tfrac{1}{2}e\pi\overline{S^2})^{\frac{1}{2}} \qquad (10)$$

a formula also obtained by Shannon(3), though our notations are somewhat different.

Apart from the factor k the entropy is also the measure of the quantity of information, in accordance with its definition as the logarithm of the number of possible, equally probable selections. The connexion between entropy is the thermodynamical meaning of the word, and information has been cleared up in 1929 by Szilárd(13) who proved the remarkable theorem that information corresponding to an s-fold selection enables the receiving system to reduce the entropy of the transmitter by a maximum of k log s. He proved also that any mechanism acquiring this information must increase the entropy by a minimum of k log s, in accordance with the second principle.

So far the concepts of information and of entropy are closely parallel, even identical. But a somewhat new feature was introduced by Shannon(5), who defined the mean entropy per symbol in a transmission in which the symbols have probabilities, i.e. relative frequencies p_i as

$$H = -k\sum p_i \log p_i \qquad H \geq 0 \qquad (11)$$

This new concept is at an appreciable remove from the physical entropy previously discussed. The probabilities p_i have no direct relation to the structure of the signal, they are determined by the source. A symbol can be represented by any configuration, or group of configurations in a basic group of cells, provided that the basic group allows at least as many distinguishable configurations as there are symbols, and even this fairly general definition does not exhaust the almost unlimited possibilities of coding. Hence the interesting properties of the expression 11. demonstrated by Shannon, must be attributed to its mathematical form rather than to its intrinsic relation with the physical concept going by the same name.

(4) THE CONNEXION BETWEEN QUANTOM REPRESENTATION AND CLASSICAL DESCRIPTION.

Comparing the representation of a signal, illustrated in Fig. 3, with the classical, mathematical description, it appears that we have lost something. Previously we had two data per cell, we are left now

with one only, a function of the quantum number N, which corresponds to the energy. What has happened to the phase?

It is not difficult to give an answer to this question on general lines. N is not an exact datum but a stochastic number, which represents a finite amount of information. But between one datum of finite accuracy and two there is no transcendental abyss of the kind which exists between a simple and a twofold infinity. One finite datum can very well contain two independent, finite data, provided that their aggregate information is not more than the original. It will be shown that this is indeed the case, and that the maximum amount of information on amplitude and phase are both contained in the single photon scale.

An electromagnetic signal can be physically analysed in various ways, but it will be instructive to consider first extreme cases only. One extreme is a counter, an instrument which records single photons. It follows immediately from Heisenberg's uncertainty principle that the "time resolution" of such an instrument canot be better than a whole cycle, hence the phase remains entirely unobservable. The other extreme is any classical field-measuring instrument, capable of recording the electromagnetic field in the signal as a function of time. It may be called for brevity a proportional amplifier, as amplification of weak signals is one of its essential functions. There is no need to consider intermediate instruments, as it will be seen that in a certain range of very weak signals every classical amplifier operates as a sort of "proportional counter".

One limit for the operation of any such instrument is given by the well known uncertainty relation of the quantum theory of radiation, (cf. Heitler,(14) p. 68)

$$\Delta N \Delta \phi \geq 1 \qquad (12)$$

where ΔN is the uncertainty in the photon number N, and $\Delta\phi$ the uncertainty in the phase ϕ. But this gives merely an upper limit in our case, as we want to determine amplitude and phase simultaneously, each as accurately as possible. It will be seen that in fact the limiting accuracy is much below what might be expected from the inequality 12.

We will approach the problem by a detailed analysis of a particular type of proportional detector-amplifier. Subsequently we will try to improve its performance to the extreme limit. It will then be found that all special features of the device vanish from the formulas, thus the final result can be considered as of general validity.

Assume that the electromagnetic signal is introduced into a rectangular wave guide, (Fig. 5) in the T_{01} mode, i.e. with an electric field in the x-direction, independent of the x-coordinate, which is at right angles to the direction of propagation z. An electron beam passes through two small holes along the x-axis, in the plane of maximum intensity, with velocity v and current intensity J. The alternating acceleration and retardation of the electrons produces velocity modulation in the beam, which by well known methods can be translated into current modulation, i.e. producing an alternating component superimposed on the mean value J. One method utilizes the "bunching" of the electrons which takes place at a certain distance from the guide, but one can also deflect the beam by a constant field and make it play between two collecting electrodes, close together. In either case one obtains an alternating current which in first approximation is proportional to the d.c. current and to the relative accelerations and retardations suffered by the electrons in the wave guide.

Let us measure the energy exchange between the electrons and the field in quantum units, $h\nu$, where ν is the mean frequency of the signal. (It will soon be seen that in order to make the exchange intense the waveband of the signal must be made so narrow that it is permissible to take the arithmetic mean.) Let $\bar{N}$ be the mean number of photons in

an information cell, which passes through a cross section of a guide in a time $\Delta t = 1/\Delta\nu$. During this time $\bar{M} = J\Delta t/e$ electrons pass through it, and exchange in the mean $\bar{n}$ quanta with the field, either by losing or by gaining energy. The positive number $\bar{n}$ is the essential parameter of the process. If $\bar{n}$ is a large number, which is possible only if $\bar{N}$ is also large, the interchange will be essentially classic, if $\bar{n}$ is small quantum phenomena will dominate.[*]

The detailed calculations may be found in Ref. 16, only the results will be discussed here. The first contains the classical theory of exchange, valid for large $\bar{n}$ and $\bar{N}$, the second a wave-mechanical calculation valid for small exchange parameters. In the classical theory the result is

$$\bar{n}^2 = \frac{32}{\pi^3}\left(\frac{2\pi e^2}{hc}\right)\frac{\sin^2\Theta}{\Theta}\frac{v}{\nu b}\left[1-\left(\frac{c}{2b\nu}\right)^2\right]^{-\frac{1}{2}}\frac{\Delta\nu}{\nu}\bar{N} \qquad (13)$$

Thus the exchange parameter is proportional to the square root of the phaton number, as may be expected. Of the dimensions of the waveguide, the width b appears explicitly in the factor of $\bar{N}$, while the depth a is contained in the angle

$$\Theta = \pi\nu a/v \qquad (14)$$

which is one-half of the "transit angle". c is the velocity of light. In addition there appears a factor, which is the reciprocal of

$$\frac{hc}{2\pi e^2} = 137$$

the fundamental number which connects photons with electrons.

Anticipating that the best measurement will require an intense interchange, we now try to increase the coefficient of $\bar{N}$ in eq. 13 by all available means. First we make the factor $\sin^2\Theta/\Theta$ a maximum. This is 0.723 and is obtained with $\Theta = 67^\circ$, i.e. at a transit angle of 134°. This disposes of the depth a of the waveguide. The optimum width b is determined by the condition that the group velocity

$$U = c\left[1-\left(\frac{c}{2b\nu}\right)^2\right]^{\frac{1}{2}}$$

must be as small as possible. But the smallest value is reached when U vanishes at the low-frequency limit, $\nu - \frac{1}{2}\Delta\nu$ of the band. Substituting these values into eq. 14 one obtains in the optimum case

$$\bar{n}^2 = \frac{1.5}{137}\left(\frac{v}{c}\right)\sqrt{\frac{\Delta\nu}{\nu}}\bar{N} \qquad (15)$$

All special features of the device have vanished in this formula, apart, perhaps, from the unimportant factor 1.5. But it is evident that the factor of $\bar{N}$ must be always much smaller than unity, while its best value,

---

[*] Quantal energy exchange between electrons and the field in a wave guide at high quantum numbers has been previously discussed by Lloyd P. Smith, (15), but we cannot agree with most of his results. Monokinetic electrons and exchange of sharply defined quanta on the one hand, well defined entrance phases and short transit times on the other are mutually exclusive phenomena by the Uncertainty Principle, hence we believe that only certain averages over Smith's detailed results have physical significance.

as will be shown later, is just unity. There exists, however, a further possibility for improving its performance. Assume that we can make each electron perform repeated passages through the guide, each transit, in the opposite direction exactly half a cycle after the last. (This is possible in principle, as the optimum transit angle is about 134°.) If the frequency were known beforehand, the number of passages would be limited only by the consideration that by repeated gains or losses the electron would be bound to fall out of synchronism. But if the signal had a single frequency, known in advance, there would be of course no communication. However, even if the frequency is known beforehand only within $\nu \pm \frac{1}{2}\Delta\nu$, one can make the number P of passages as great as $\nu/\Delta\nu$, without risking a phase error of more than $\pm \frac{1}{2}\pi$, and it can be shown that these passages are almost of equal value, so that $\bar{n}$ is increased very nearly by a factor P. The number of passages required to make the coefficient of $\bar{N}$ in eq. 15 unity is

$$P_{opt} = 9.5\left(\frac{c}{v}\right)^{\frac{1}{2}}\left(\frac{\nu}{\Delta\nu}\right)^{\frac{1}{4}} \qquad (16)$$

i.e. at least of the order ten. But this number must not exceed $\nu/\Delta\nu$, hence we obtain the condition that in order to realise optimum conditions the frequency band must be so restricted that

$$\left(\frac{\nu}{\Delta\nu}\right) \geq 20{\cdot}2\left(\frac{c}{v}\right)^{2/3} \qquad (17)$$

This is a somewhat surprising result. In the mathematical representation it did not matter whether we divided up the frequency band into broad or narrow strips. But by the intervention of the number 137, ($20.2 = (137/1.5)^{2/3}$), it turns out that only narrow frequency bands are capable of accurate analysis by means of electrons. (If ions with charge Z were used one would have to replace 137 by $137/Z^2$, and the condition would be less stringent.)

Thus, at least in theory, the device could be perfected up to the optimum performance. The practical difficulties are of course evident. In practice one would rather replace the wave guide by a "high-Q" resonator, but this would somewhat complicate the theory.

We thus find, assuming P passages, in the "classical" case

$$\left(\frac{\bar{n}}{P}\right)^2 = \frac{1.5}{137}\left(\frac{v}{c}\right)\sqrt{\frac{\Delta\nu}{\nu}}\,\bar{N} \qquad (18)$$

while the corresponding wave-mechanical formula, valid for very weak interchange is,

$$\frac{\bar{n}}{P} = \frac{1.0}{137}\left(\frac{v}{c}\right)\sqrt{\frac{\Delta\nu}{\nu}}\,\bar{N} \qquad (19)$$

Apart from the factor 2/3 the coefficient of $\bar{N}$ is the same in both cases, but this time the quantum exchange $\bar{n}$ is proportional to the photon number itself, not to its square root. It can be said therefore that for small photon concentrations the device acts as a counter, at large concentration as a field measuring instrument. The intermediate region is difficult to calculate, but as shown in Fig. 7, the two branches can be connected by a plausible curve.

One can also interpret the results in this way: If there are few photons present, there will be few collisions, and equal probabilities of gains and losses, at any instant. With increasing photon concentration repeated collisions will increase in number, and the resulting loss or gain increases with the square root of the photon number only; but this

resultant has now a prevailing direction, which changes its sign with the frequency of the signal. At this stage the "classical field" has developed.

Having ascertained that, with certain reservations, we can make the exchange as strong as we like, we ask the question:- If we know the average photon number $\bar{N}$, how must we adjust the electron beam current J, and the exchange parameter $\bar{n}$ in order to measure the field amplitude E with maximum accuracy? And having made these adjustments, how many steps shall we be able to distinguish in the scale of the field amplitudes? - Evidently this question has a precise meaning only in the "classical" range of large $\bar{n}$ and $\bar{N}$, and the following considerations relate only to this case. In order to simplify the problem we neglect the thermal noise, i.e. we put $N_T = 0$, so that the relative accuracy on the photon scale would be $1/\bar{N}^{\frac{1}{2}}$, and the total number of steps in the photon ladder $2\bar{N}^{\frac{1}{2}}$. The calculations are carried out in Ref. 16, here we give only the physical considerations.

The quantity to be measured is the electric amplitude in the information cell, which is proportional to the square root of the photon number. The measured quantity, on the other hand, is the alternating electron current, which, as mentioned above, is proportional to $\bar{n}.J$ or to $\bar{n}.\bar{M}$ for not too strong signals, $\bar{M}$ being the mean number of exploring electrons per cell. For the optimum we impose the condition that the relative mean square deviation of the quantity measured from the quantity to be measured shall be as small as possible, i.e.

$$\frac{\overline{(nM - C\bar{N}^{\frac{1}{2}})^2}}{(\bar{n}\bar{M})^2} = \min. \qquad (20)$$

where the proportionality factor C is determined from the condition

$$\overline{(nM - C\bar{N}^{\frac{1}{2}})} = 0$$

We have to choose $\bar{n}$ and $\bar{M}$ so as to satisfy the condition 20. for given $\bar{N}$. There must be an optimum, for these reasons: A too weak interchange $\bar{n}$ would leave the cell unexplored. A too strong interchange on the other hand will interfere with the object of the measurement and spoil it. Though in the mean electrons are as often accelerated as retarded, fluctuations in the numbers M_1 and M_2 of electrons which pass through accelerating and retarding phases might produce extra photons, which could not be distinguished from those belonging to the signal, or annihilate some. The spurious photons are generated according to a law

$$\delta_2 N = \bar{n}\,(M_1 - M_2) = \bar{n}\,(\delta M_1 - \delta M_2) \qquad (21)$$

as $\bar{M}_1 = \bar{M}_2 = \frac{1}{2}\bar{M}$. We have written $\delta_2 N$ for this number, considering it as a fluctuation which must be added to the natural fluctuation $\delta_1 N$, whose law is $\delta_1 N^2 = \bar{N}$. The two fluctuations must be considered as independent.

It is already evident from the above that the fluctuations in the number of beam electrons i.e. the "shot effect" plays an important part in these phenomena. A too weak current has a high relative fluctuation. A too strong current, especially aided by a large exchange will again spoil the object. It may be noted that we have here a type of uncertainty which springs directly from the fact that photons and electrons are discrete, without any reference to the physical values of h and of e.

The relative error according to eq. 20 is calculated on the basis of eq. 21, together with such evident assumptions as the independence of the fluctuations of n and M and the "natural" part of δN. We assume also "normal" shot effect, $\delta M^2 = \bar{M}$. The result is

$$\frac{\overline{(nM - CN^{\frac{1}{2}})^2}}{(\overline{nM})^2} = \frac{1}{4\overline{N}} + \frac{1}{\overline{n}} + \overline{n}\left(\frac{1}{\overline{nM}} + \frac{\overline{nM}}{4\overline{N}^2}\right) \qquad (22)$$

This is a minimum for

$$\overline{nM} = \overline{n}.\overline{M} = 2\overline{N} \quad \overline{n}^2 = \overline{N} \qquad (23)$$

from which $\overline{M} = 2\,\overline{N}^{\frac{1}{2}}$. This gives the simple rule that for optimum analysis of the signal one must take one electron for every step in the scale of the photons, and the interchange $\overline{n}$ must be itself equal to one distinguishable step at the level $\overline{N}$. This again is a general rule, quite independent of the special model from which we started.

Substituting these values, we obtain for the minimum of the mean square relative error in the measurement of amplitudes

$$\left(\frac{\overline{(nM - CN^{\frac{1}{2}})^2}}{(\overline{nM})^2}\right)_{min} = \frac{1}{4\overline{N}} + \frac{2}{\overline{n}} = \frac{1}{4\overline{N}} + \frac{2}{\overline{N}^{\frac{1}{2}}} \qquad (24)$$

As we are dealing with large photon numbers only, the first term at the right hand side can be neglected with respect to the second. Thus we see, applying the same rule which we have used in constructing the photon ladder, that the smallest distinguishable relative step in the amplitude scale is $\sqrt{2}$ times larger than the square root of the corresponding quantity in the photon scale. In other words, the proper scale of the amplitudes will contain always less than the square root of the number of steps in the photon scale, apart from the factor $\sqrt{2}$ for the reason that the optimum setting is of course possible for one level only.

Having determined the proper scale of the amplitudes, a simple application of the Uncertainty Principle, shows that the proper scale of the phase must also contain less steps than the square root of the photon scale. Thus, summing up, we see that the classical description of the signal, by being too detailed, gives in fact a somewhat smaller total amount of information than the quantum description.

The classical method of description, though theoretically inferior, may of course be still the best practically in the range of frequencies used for electrical communications, where efficient photon counters are not available. Conditions are different in the optical region, where detectors of the counter type - such as the eye - are not far from perfection. In this region analysis in terms of electromagnetic waves is as yet technically impossible, but it is interesting to note that even if it were possible, it would not be very practical. Progress in the field of microwaves is now actually approaching a region where the two different methods of analysis may become competitive. We see from our results that it takes about a hundred million photons per information cell in order to define amplitude and phase of the signal to 1% each. Remembering that at 1 cm wavelength the number of thermal photons per cell is only about 200, it may be seen that the time may be not far off when the imperfections of the classical method of description will manifest themselves even in electrical communications.

It may be hoped that these considerations have shown that the concepts of information theory may well prove their usefulness when applied to problems of physics.

REFERENCES

1. Gabor, D. Journ. I.E.E. 93, III, 429-457, 1946, ibid; 94, III, 369-387, 1947; Nature, 159, 591-594, 1947.

2. Ville, J.A. Cables et Transmission, 1, 61-74, 1948.

3. Shannon, C.E. Proc. I.R.E. 10-21, 1949.

4. Tuller, W.G. Proc. I.R.E. 468-78, 1949.

5. Shannon, C.E. loc. cit. and Bell Syst. Techn. Journ. 27, 379-423, 623-656, 1948.

6. Oswald, J. Comptes Bandus, 229, 21-22, 1949.

7. v. Laue, M. Ann. d. Phys. 4 44, 1197-1212, 1914.

8. Born, Max "Natural Philosophy of Cause and Chance", Oxford, 1949 p. 81.

9. Nyquist, H. Phys. Rev. 32, 110-113, 1928.

10. Szillárd, L. Zeitschr. f. Phys. 32, 753.788, 1925.

11. Mackay, D.M. Phil. Mag. (7) XLI, 289-311, 1950.

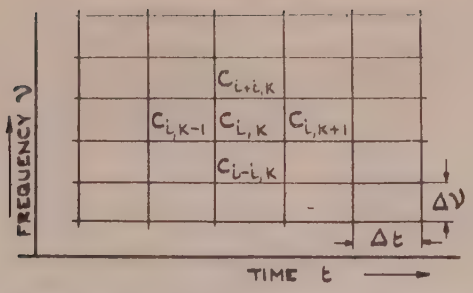
12. Jordan, P. "Statistische Mechanik auf quantentheoretischer Grundlage" Fr. Vieweg, Braunschweig, 1933.

13. Szilard, L. Zeitschr. f. Phys. 53, 840-856, 1929.

14. Heitler, W. "The Quantum Theory of Radiation", Oxford, 2nd Ed. 1944.

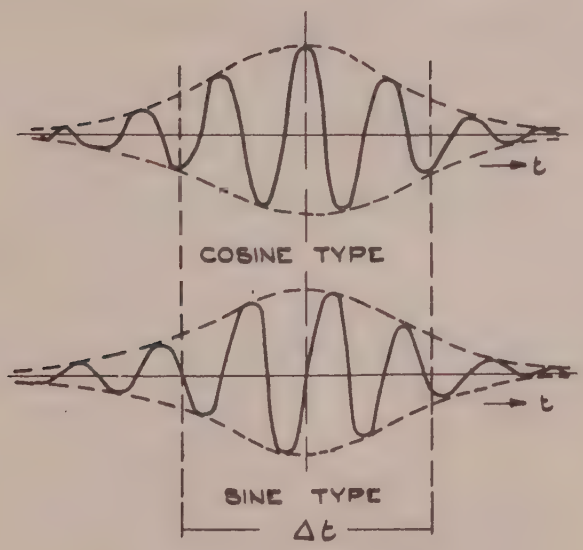
15. Lloyd P. Smith, Phys. Rev. 69, 195-210, 1946.

16. D. Gabor. Phil. Mag. (7), 41, p. 1161, 1950.



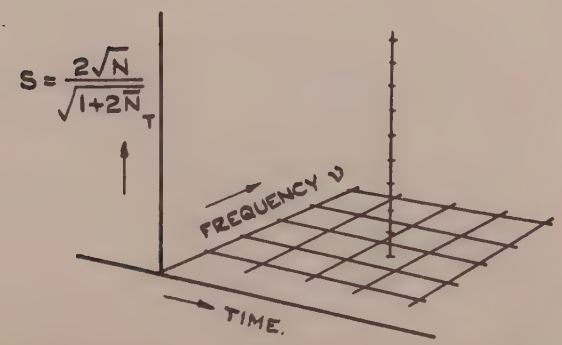
INFORMATION DIAGRAM. THE TIME-FREQUENCY HALF-PLANE IS DIVIDED UP INTO CELLS OF UNIT AREA AND AN ELEMENTARY SIGNAL IS ASSOCIATED WITH EACH, WITH A CO-EFFICIENT C_{iK}.

FIG. 2.



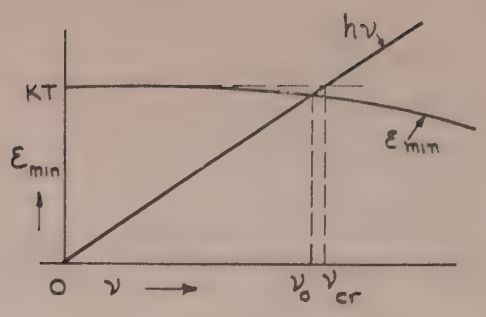
ELEMENTARY SIGNALS OF THE "COSINE TYPE", (EVEN) AND OF THE "SINE TYPE" (ODD)

FIG. 3.



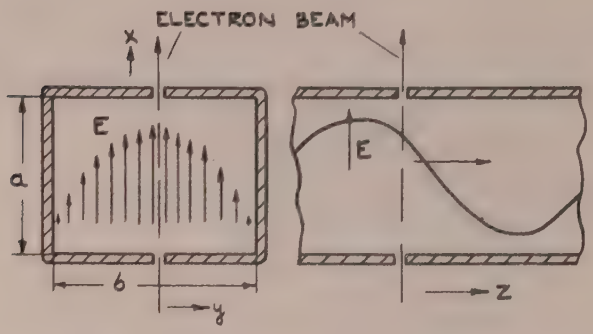
REPRESENTATION OF AN ARBITRARY SIGNAL. A LADDER OR "PROPER SCALE" OF DISTINGUISHABLE VALUES IS ERECTED IN EVERY CELL, ON WHICH THE OCCUPATION IS MARKED OFF.

FIG. 4.



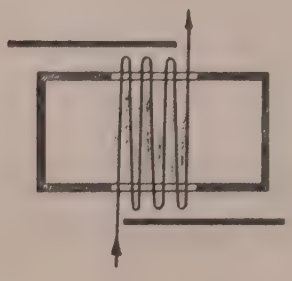
THE ENERGY REQUIRED FOR THE TRANSMISSION OF THE FIRST "BIT" OF INFORMATION.

FIG. 5.



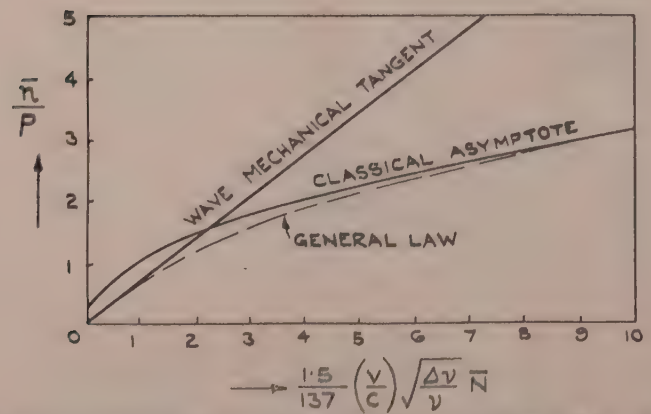
ANALYSIS OF AN ELECTROMAGNETIC SIGNAL IN A WAVEGUIDE BY AN ELECTRON BEAM.

FIG. 6.



REPEATED PASSAGES.

FIG. 7.



MEAN QUANTUM EXCHANGE BETWEEN ELECTRONS AND PHOTONS. THE INITIAL TANGENT IS CALCULATED FROM WAVE MECHANICS. THE ASYMPTOTES

QUANTAL ASPECTS OF SCIENTIFIC INFORMATION

by
D. M. MacKay

SUMMARY

This paper is an attempt to clarify some aspects of the approach to experimentation suggested by the author in a recent publication[4] in the Philosophical Magazine, (hereafter referred to as P.M.). The concept of "amount of information" is shown to have three distinct senses in current literature. Two of these are definable as numerical features of the logical pattern of propositional relations which we make to represent a result. The third measures the relative unexpectedness of the pattern, which may or may not be connected with its numerical features. Since logical patterns can be built up from discrete quantal elements, the information enabling them to be built is quantised by our use of 'yes-or-no' logical forms as scientific statements. The number of discrete 'elementary propositions' in a given description cannot be altered by any complete reformulation. This is seen to be the basis of our ability to 'barter' certain quantities for one another - e.g. accuracy for speed of response in a galvanometer or a communication-channel.

The term selective information is suggested to distinguish the third sense of 'information' (called amount of detail in P.M.) from the first two, which measure respectively the number of independent features (structural information) and the weight of evidence or precision (metrical information) in a result. 'Selective information' is the measure of information currently used by communication engineers, and a distinguishing title appears to be essential.

From the standpoint here adopted the various uncertainty-relations of physics illustrate a general axiom expressing the quantal nature of the logical descriptions which we make. Some other practical and theoretical implications of the theory are examined.

(1.) FUNDAMENTAL NOTIONS

(1.1) 'Information' in everyday speech.

In everyday language we say we have received information, when we know something now that we did not know before. If we are exceptionally honest or a philosopher, we assert only that we now believe something to be the case which we did not previously believe to be the case. The word information is thus used commonly in several senses, corresponding to the several kinds of change which can take place in our knowledge.

Three senses in particular interest us as scientists, and are typified by the following three sentences:

Three senses of Information

(a) "I got little information from what he said; I knew pretty well what he'd say before he opened his mouth".

(b) "An instrument which can respond in 1/100 sec. gives me more information per second than one which can respond in 1/10 sec."

(c) "An experiment yielding a result accurate to 1% gives me more information than one accurate to 10%.

Each of these refers to a different way in which our knowledge has increased. Roughly speaking, the first considers the unexpectedness of the addition; and the

second its complexity of form; and the third the degree of confidence we have in it. We shall meet each of these in due course, in more precisely defined contexts.

(1.2) Information as something measurable.

A hundred years ago it would have been thought merely a pardonable error if a stranger were to suppose 'information' to mean something measurable, like energy for example. "Amount of information", he would be told, is a metaphorical term, and has no numerical properties.

But in fact he would have been right, though in grave danger of utter confusion unless he learned to distinguish between the complementary senses we have met. It is to be hoped that he would do so, rather than adopt the Procrustean method and insist on 'refuting' all except one definition.

We might profitably compare our present position with that of a man who has never heard of the concept of 'size'. He discovers that in at least one sense of the term, a man, a sack of potatoes, an oil-drum, and a tree-stump can all be said to have "the same size". He has now made some progress. At least, he knows that 'size' can not mean any of the properties which are not common to all four examples. With patience, he might eventually arrive at the notion of size by a process of elimination. He might even discover the different senses of the term, and coin the terms 'volumetric size' or 'volume', 'superficial size' or 'area', and 'linear size' or 'length', to distinguish three of its important senses. (He would doubtless meet with opposition from some of his colleagues at this point, and be accused of making things difficult with jargon for which they saw no need; but the parable need not be developed!)

But he has also a second line of approach which is more fruitful. He asks us: "What differences does size make in an object?" "In what circumstances do you become aware of it?" "To what does size make a difference?"

He discovers that we define the size (volume, area, etc.) of a body in terms of its ability to cause changes of a certain kind, in certain circumstances (for example, the indication of a point on a scale, as a result of certain well-defined operations with appropriate measuring apparatus.) This is his clue. By systematically studying what happens when something is affected by the size of an object, he discovers the meaning of the term.

Our task is a similar one. We shall find it profitable to ask: To what does information make a difference? What are its effects? This will lead us to an 'operational' definition covering all senses of the term, which we can then examine in detail for measurable properties.

(1.3) Representations

Information makes a difference to what we believe to be the case. Its effect is to change, in one way or another, the total of "all that is the case" for us. This rather obvious statement is the key to all definitions of Information. Any aura of metaphysical abstruseness which it may have is easily exorcised. What we know or believe, in science at least, could be represented in a variety of quite precise ways; we might make a long statement, or draw a

symbolic picture, or make a physical model, or send a communication-signal. All the results could in a sense show or 'contain' what we know or believe to be the case.

Representations

All are examples of what we may call representations: structures which have at least some abstract features in common with something else which they purport to represent. It is these abstract features of representations which we want to isolate. They form the real currency of scientific intercourse, which is normally obscured in wrappings of adventitious detail.

Now that we have established this fundamental notion of a representation, Information can be described as what we depend on for making statements or other representations.

Information in General

More precisely, we may define information in general as that which adds to a representation of what is known or believed or alleged to be the case.

(1.4) Criteria of "Amount of Information"

We have already seen in paragraph 1 that there are different criteria by which "the amount of information" given by a representation can be estimated. Two quite different approaches are possible.

(1.4.1) The first pays no attention to the structure of the representation per se. Its criterion is simply the

Unexpectedness

unexpectedness of the representation. This need not have any simple relation to its form. A long message which is frequently received gives less information, in this sense, that a short one which is more uncommon.

This is (particularly though not exclusively) the communication engineer's sense of the term (5,6,7,8) and will be fully dealt with by other speakers in this symposium. To measure the amount of information given by a message, he considers it as if it had been selected by some standard procedure from an assembly of possible alternatives. The

Assembly Ensemble

assembly (or 'ensemble') is made up of all possible messages, in the proportions in which they are expected. Thus, if the most efficient selection-procedure is used, the commonest messages can be selected in the smallest number of steps, and the complexity of the corresponding selection-process - the number of yes-or-no choices determined - becomes the criterion of the amount of information given by any particular message.

The details need not now be considered (they have been summarised briefly by the author in the Glossary). We may,

Selective Information

however, suggest the name Selective Information as an essential means of distinguishing this measure from the others which are our first concern. Selective-information-content is then a measurable property of a selection-process rather than of a representation. It has clearly nothing whatever to do with the meaning of the representation.

(1.4.2) The second approach is appropriate when the abstract features of the representation itself are the centre of interest. This is in the main the approach of Scientific Information-theory, which seeks to make abstract representations of what has happened, and does not initially consider expectedness or unexpectedness when discussing the amount of information contained in a representation.

Information-content in this sense, (which we have seen in paragraph 1 to have two aspects), is a measurable property of a representation. It is this concept with which we shall first be concerned.

(2.) MEASURABLE PROPERTIES OF REPRESENTATIONS

(2.1) Sentences

It is perhaps not obvious that a change in "what is the case" can be measured, as a change in a size-measuring system can. But in science at least our processes of acquiring information are designed to enable all the information gained from an experiment to be represented (in principle) in sentences. So if we can find a way of analysing sentences quantitatively, we should be able to measure information in terms of its effects on sentences or logical propositions representing "that which is the case" as we know it.

We shall quite deliberately confine ourselves to discussing knowledge which can be completely expressed in such 'measurable propositions', of the sort considered in mathematical logic. This begs no questions as to the validity of other forms of knowledge. But scientific propositions are particularly designed so that they can be verified or contradicted in a 'yes-or-no' fashion; and this places them well within the mathematical logician's field - in so far as they conform to their professed ideal. Our donning of blinkers therefore need cause us no alarm. The only danger indeed is that we may be so little aware of their presence as to forget their limitations. They have, however, the popular merit of protecting us from the currently unforgiveable sin of "committing metaphysics".

(2.2) Elementary propositions

Atomic Propositions

The logician's way of measuring a statement is to break it down into a set of simpler statements. If this process is carried to its logical conclusion, we are left with a number of simple assertions, so elementary that they cannot be analysed further. Wittgenstein[9] has called these simplest elements atomic propositions. An atomic proposition is so simple that it raises only two possibilities - that its simple assertion is true or false. Conceptually, it defines just two alternatives.

The Information Pattern

Each atomic proposition asserts that some simple relationship holds between two entities. A set of such propositions can then be thought of as describing a logical pattern of relationships. The business of a sentence, or any other representation, is to convey this pattern of relationships to another mind. It is this pattern which constitutes the irreducible common feature of all representations which we should describe as 'equivalent'. This is in fact the information-pattern for which we have been looking.

We may now think of an atomic proposition as asserting the existence of one element in the information-pattern to be conveyed, or alternatively as an instruction to the receiving mind to add one element to the pattern which the statement is enabling it to reproduce. Each element in the pattern is so simple that its only property is existence. (i.e.) It appears in the pattern if the corresponding atomic proposition is true;

it does not appear if it is false. Each element therefore owes its individuality only to its position in the pattern, and may be regarded as a kind of link between two [illegible] in the pattern.

Any pattern representing experience must ultimately make contact with elements representing the primitive sense-data in terms of which it acquires meaning; but this aspect does not concern us here. We need merely note in passing that in practice every ordinary word we use is equivalent to an information-pattern corresponding to the relationships by which it was defined and has meaning for us. Forgetfulness of this fact is believed to be the source of much confusion in discussions on probability.

(2.3) The quantization of Scientific Information

Scientific Statements

A scientific statement is ideally based entirely on observable evidence. It may be thought of as a set of instructions enabling the reader to make for himself a representation of the observer's experience. The ideal statement which would contain all the information supplied by a particular experiment can presumably be dissected into elementary propositions relating to observation. In this way the ideal description given is irreducibly broken up or quantized into discrete propositions, each of which asserts the occurrence of one elementary event, which may be the observation of a new (i.e. previously unknown) relationship either in time or space.

Quantization

This discreteness is forced on us by our use of logical forms - by our apparent inability to communicate our experience unambiguously in any other way. To speak of an 'imperceptible addition' to a logical form is meaningless. We can therefore define a unit of information for each sense of the term 'information' that we may encounter, as that which decides us to make the minimal addition possible, of the corresponding kind.

Units of Information

It might perhaps be argued that the continuous field of real numbers enables us to make changes as small as we wish in a statement. If the statement represents experience, however, we must be prepared to justify any change by reference to our experience. The statement we should then have to make would itself be quantal, if precise, and would represent the additional information gained. To speak of an imperceptible addition to our experience is self-contradictory.

This does not mean that continuously variable quantities will not be found appropriate to represent aspects of "the case" as we believe it to be. It means merely that the activity of coding experience into scientific statements is a quantal one, and leads to irreducibly quantal descriptions of experience.

(3.) SCIENTIFIC INFORMATION

(3.1) The two prerequisites for the making of Scientific statements

A scientific statement is a precise record of identifiable events. The last two words bring up two essential and complementary requirements which must be met before a scientific statement can be made.

(3.1.1) The a priori Contribution

In the first place, we must have the means of classifying - identifying, locating, labelling, distinguishing - the events we hope to be able to describe. Our experiment must be prepared in such a way that each event as it occurs finds a prepared and distinct category into which it fits and as a member of which it can be identified and spoken of. For example, we cannot say when events occur, unless we have some means of logically separating out (pointing out) distinct instants or rather intervals of time with which they can be 'labelled'.

This is an a priori feature of our statement of the result - prior, that is, to the making of the particular experiment. It may itself have originally involved an experiment to determine the number of such categories provided by given apparatus. But once determined, this knowledge counts as prior information in subsequent applications of the apparatus. For example, an experiment is usually needed to establish the response-time of a galvanometer. But once it is known, we can calculate the number of practically independent readings per minute which the instrument can give, and this figure and its implications counts as prior information in any subsequent experiment using it. We can think of the instrument as enabling us meaningfully to speak of just this number of points on a time-scale covering one minute, and no more. These then are the time-labels which we may use in making our statement.

Structural Information (see below)

We may call such prior information "structural information"; the term will be more precisely defined below.

(3.1.2) The a posteriori Contribution

In the second place, we must perform the experiment, and observe events. Now we normally observe events which lead us to inferences about physical quantities in which we are interested. We observe a pointer-reading, and infer the magnitude of a current. We see a fluctuating trace on an oscilloscope screen, and infer the magnitude of a voltage. We express the result of each reading by a number having a certain standard error. We label (identify) this number by means of the appropriate one of our distinguishable categories, and it can then appear in the statement we are making.

Metrical Information (see below)

The second, a posteriori, contribution to our statement is therefore made by the occurrence of observed events. The more we observe, in general terms, the greater the weight of evidence to which our statement is equivalent - the greater the weight which if we make it properly we may say it contains. This weight of evidence, suitably defined, we shall call the amount of "metrical information" yielded by an experimental measurement. The performance of the experiment then puts (b) the raw material of actual experience into (a) the prepared framework of categories. These two features together form the necessary ingredients of a scientific statement.

It may perhaps be objected that if a statement of experience cannot be made without both structural and metrical features, then to speak of structural information apart from metrical information is meaningless. This

however would be to misunderstand the terms. The number of independent coefficients required to specify a given signal for example depends only on the frequency-characteristics of the channel through which it comes, and is known a priori. The precision with which these may be specified depends inter alia on the strength of the signal, and can only be determined a posteriori. It is true that any signal we describe scientifically must have happened in order that we can say something about its coefficients or degrees of freedom. But there is a perfectly clear distinction between their number (known beforehand) and their individual weight or reliability (known only afterwards).

(3.2) Structural Information

(3.2.1) The unit of structural information

We say that one piece of apparatus gives us more structural information than another, when there are more independently-variable features or degrees of freedom in the logical description we can make of the result. A microscope which can distinguish points 10^{-4} cms apart at the limit of resolution gives us more structural information per square centimetre than one limited to 10^{-3} cms, because in describing a given area of specimen we can formulate a greater number of independent propositions about the intensity of light as a function of position. A communication channel which has a wider bandwidth than another, gives us more structural information per second, because in describing a signal occupying a given time-period it enables us to specify a larger number of coefficients defining the amplitude as a function of time.[2] In each case as we have seen, the term structural information connotes "that which enables us to formulate independent propositions" about the function in which we are interested.

We can therefore define a unit of structural information as that which enables us to prepare one independent category for one independent feature of the result. For this unit there already exists the suitable name of a logon, defined in a less general sense by Gabor[2] for the case of communication-signals. The amount of structural information in a result or its logon-content, indicates the number of independent categories or degrees of freedom precisely definable in its logical description.

Logon-content

With each of these there will normally be associated one observed result representing experience, which will correspond to a body of elementary propositions about events. The logical representation we make of the total result will thus consist of a number of distinguishable (i.e. independent) groups or clusters of elementary propositions. The number of these groups is the fundamental definition of the amount of structural information in the representation, or its logon-content. The latter term is felt to be preferable because (a) granted familiarity with the concept of a logon, it is precise and self-explanatory, (b) it avoids use of the word 'information', which we have seen must be hyphenated with an adjective in order to be usable at all unambiguously.

(3.2.2) Logon-capacity

In both the examples we have discussed, structure is defined in terms of a reference-coordinate, of space or time. In such cases we can define an important measure of

Logon-capacity

the capacity of an instrument or experimental method to give structural information. The logon-capacity is defined as the number of logons which are provided by the apparatus or method per unit of coordinate-tract (centimetre, square centimetre, second, etc.). It is really the same notion as "resolving power", defined in a more precise and general way, and may be thought of as an index of 'bandwidth' in a generalised sense.

Thus the logon-capacity of a microscope in a particular region in the focal plane can be defined in logons/cm^2, and is a measure of the optical resolving-power in that region. The logon-capacity of a galvanometer or a communication channel is measured in logons/sec., and represents the 'time-resolving power' or the number of (practically) independent readings per second which can be made with the apparatus.

The total number of independent categories or features in the result - the logon content - is then the integral of the logan-capacity over the extent of coordinate tract occupied by the result. If for example the logon-capacity of a communication channel is constant, the logon-content of a signal is simply the product of logon-capacity and duration.

(3.3) Metrical Information

(3.3.1) Precision and scale-units

Suppose that we have made a single measurement of some magnitude. How precisely can we specify the magnitude we have measured? Obviously if we can read the scale to three places of decimals, we could give three places of decimals. But if the magnitude is actually known to be fluctuating with a certain standard error, we may know that such precision would be unjustified. Accordingly if we are wise we make our statement only in a form sufficiently modest to have a reasonable chance of being verified.

Scale-unit

If we were to graduate the scale ourselves, the useful limit to the separation between marks, or the scale-unit, would be such that our next reading is (on the basis of our data) as likely to fall into the same scale-interval as before, as to fall outside it. In other words, a statement of the form "x extends from the first to the m'th interval" - i.e. as we will say, "occupies m intervals on my scale" then has a 50-50 chance of being disproved by the next observation. (We are assuming that x has a fixed mean-value,). This gives the statement the logical 'weight' of an elementary proposition, which is presumably the minimum weight justifying its appearance.

The scale-unit of our magnitude is thus in the limit equal to twice the so called 'probable error', or about 1.35 times the standard deviation if we assume a normal distribution. If we describe our result in terms of a smaller unit, we are bound to do so in a form having a probability less than one-half - more likely to be false than true. The figure of one-half implies, not of course that we are left in complete ignorance by our measurement, but that it has enabled us to divide the whole range of possible values of x into one small interval and two large remaining outside regions, and to assert that x is as likely to be in the small interval as in the large regions outside it.

There may sometimes be cases in which the fluctuation in x is not independent of its magnitude. Our scale-unit as defined for x may therefore vary in size according to the magnitude of x. There will generally however be some function of x (x^2, or $\sqrt{x}$ for example) which does have a constant fluctuation. We could therefore specify the observed magnitude of this function by stating simply the number of intervals it occupies i.e. the number of equally significant steps in its magnitude on a scale graduated in intervals (scale-units) equal to twice its probable error. Such a scale, on which equal intervals are equally significant or 'equiprobable', is termed the proper scale* for the quantity concerned. The proper scale of voltage for example in the presence of constant thermal noise, is a scale linear in voltage, graduated in intervals equal to the (constant) "probable range" of fluctuations. Larger intervals (e.g. twice the standard deviation) may of course be used, but the maximum number of significantly-distinct levels of voltage is attained using the smallest logically justifiable scale-unit of voltage.

Proper scale

(3.3.2) Metron-content of statistical samples

If we were looking for an index of precision, we might be tempted to take the number of intervals occupied on the proper scale as a suitable measure. If however we are looking for a measure of the amount of information represented by a measurement, in the sense of that which gives precision or reliability to the result, (which we later distinguish as metrical information) we must consider the matter more closely.

It is true that the number of proper-scale intervals increases as the precision increases; so that we should expect a suitable measure of information in this sense to be a monotonic increasing function of that number. But we wish to equate our measure of information to the number of elementary propositions, of some sort, subsumed in our description of the result. Now it is an axiom of logic that the number of elementary propositions into which a statement can be broken down can nover be increased by a logical reformulation of the statement, and is invariant for all reformulations which are logically complete.

If then we consider two separate and similar measurements (samples) of a magnitude, to represent a certain number of elementary propositions relating to magnitude, then the result of combining those observations must be to produce a statement containing double the number of elementary propositions. But it is well known that the standard deviation of such a combined figure is not halved, but divided by $\sqrt{2}$. It is the variance (the square of the standard deviation) which is halved. In fact the reciprocal of the variance of a statisctical sample is proportional to the size of the sample, provided that the elements of the sample are independent.

These considerations among others led R. A. Fisher[1] to define the 'amount of information' (in our metrical sense) given by a sample as the reciprocal of its variance. His

---

* This definition differs from the rather loose one given in P.M., where an error in the illustration in para. 4 (c) unfortunately obscures the issue.

precise definition is given in the explanatory glossary, and deserves fuller discussion than was given there or can be given here. But the essential point is that his "amount of information" is an additive measure. As defined, it is not dimensionless, but requires to be multiplied by a quantity having the dimensions of the variance - i.e. magnitude squared, - if it is to be independent of changes of unit.

This definition could have been the clue to the function for which we are looking. The square of the number of proper-scale intervals occupied is proportional to the reciprocal of variance, and is dimensionless. It is thus proportional to the size of the "sample" we have effectively taken by making our observation.

We may therefore think of our observation as specifying a number on a new 'conceptual scale', on which the number of intervals conceptually pictured as occupied is the square of the number occupied on the proper-scale, and measures what we have termed the amount of metrical information in our sample. We may define a unit of metrical information, or one metron, as that which raises the number of occupied conceptual scale-units by one. This in turn may be thought of as the consequence of one 'elementary event' - the arrival of a corresponding 'unit-contribution' to the sample which we have observed.*

Metron

(3.3.3) Occupance-relations and coincidence-relations

We have arrived at a representation which in suitable cases enables the number of elementary intervals occupied on a conceptual scale to be interpreted as the number of elementary propositions (relating to magnitude) yielded by a result. It seems reasonable to identify the two and take as the elementary proposition an assertion of the form "an interval is occupied". i such propositions correspond to a representation in which i intervals are occupied on the conceptual scale, or $\sqrt{i}$ on the proper scale.

Occupance-relation

Coincidence-relations

A proposition of this elementary form we say asserts an occupance-relation between a magnitude and a scale-interval. Logically it is not quite unanalysable or 'atomic', being equivalent to two assertions of coincidence-relations between the ends of the unknown interval and the ends of the scale-interval. We can think of propositions asserting these coincidence-relations as the most elementary propositions relating to observation, (i.e.) as the atomic propositions of the scientific method. The elementary metrical proposition asserting occupance is thus the lowest 'molecular' propositions, and is the simplest possible proposition relating to measurement.

(3.3.4) Metron-content

To sum up, we have seen that in the case of quantities subject to statistical fluctuations, these fluctuations provide a natural scale-unit for the dimensionless specification of the magnitude of the quantity (i.e. as a pure number). If the fluctuations are independent of the magnitude concerned, the latter may be represented on a proper-scale whose graduations are just far enough apart

* The arrival of one photon in a unit-cell in the quantum regime treated by Gabor[3] is an example.

to give a reading on the scale a probability of one-half of being confirmed. The number of metrons provided by the result, which we may term the metron-content, is the square of the number of intervals occupied on the proper scale, and may be thought of as the number of intervals occupied on an abstract conceptual scale on which each occupied interval represents one elementary event out of the total (effectively) observed.

Metron-Content

It will be noted that metron-content is necessarily positive. A reading of a negative voltage can give the same metrical information as that of a positive. This indeed could have suggested the square of the proper-scale reading as more suitable than the reading itself to be a measure of metrical information-content.

It may also be observed that if for example we were measuring power instead of voltage, then the proper scale would still by definition be that of voltage if the noise-voltage were independent of the power level. The scale of power is thus the scale also of metron-content. Equal increments in power correspond to equal increments in metron-content for a given noise-level. The physical interpretation here is illuminating. Doubling the power doubles the number of units of energy which have been contributed to the sample in a given time. The 'elementary events' here are the arrival of elementary units of energy, equal in magnitude to the 'noise-energy' in the sample, or the noise-power per (logon per second).

(3.3.5) Metron-content and entropy

We can follow this point a stage further. Thermal noise-energy per degree of freedom is directly proportional to absolute temperature, in the classical case. (This fact is well known to communication engineers in the form of Nyquist's expression for the mean-square fluctuation or variance of voltage across a resistor R at temperature T, measured through a transducer with rectangular bandwidth Δf:

$$\overline{\Delta V^2} = 4kTR\,\Delta f. \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Here k is Boltzmann's constant.) Evidently then what determines the number of elementary propositions $V^2/\overline{\Delta V^2}$ in the classical case, is not the total energy per sample per se but as the author has elsewhere[4] pointed out, the ratio of this to the absolute temperature. This quantity $\left[V^2/4kR\Delta f\right]/T$ has the dimensions of thermodynamic entropy, and represents the minimal amount by which the thermodynamic entropy of the system observed must increase, in order that the measurement may be described with the observed precision[*]. In other words each elementary occupance-relation costs for its establishment a certain minimal increase of the entropy of the total environment.

Entropy

It may be of interest to calculate the size of this somewhat artificial 'quantum' of entropy, for the case treated by Nyquist. It is known[2] and can be deduced from purely logical considerations[4], that the number of logons yielded on one second by a transducer with bandwidth Δf is $2\,\Delta f$. The structural scale-unit of time, during

---

[*] This must not be identified with the entropy of a selection, introduced by Shannon; The connection is not obvious, and will be examined in the author's second paper.

which one independent measurement can be made, is therefore $\Delta t = 1/2 \Delta f$. If under optimum conditions a generator with internal impedance R is supplying a load resistance R with a voltage V, the mean square noise voltage ΔV^2 will be one quarter of Nyquist's figure, which would apply on open-circuit. The proper-scale of voltage will then be graduated (with our present convention) in intervals of magnitude $1.35 \sqrt{kTR\Delta f}$. A voltage V will therefore occupy $V/1.35 \sqrt{kTR\Delta f}$ intervals, and will correspond to a metron-content i of $i \doteqdot V^2/1.82kTR\Delta f$. Since V^2/R is the power W received, the energy received per logon is $(V^2/R)\Delta t$ or $(V^2/2R\Delta f)$. The corresponding entropy-increase Δs must be not less than $(V^2/2RT\Delta f)$, which is (0.91k) times the number of metrons.* Each metron costs at least 0.91k units of entropy.

(3.4) Representations of magnitude in general

The present discussion has only touched the fringe of the question of representing magnitude in general in an abstract pattern. Much work remains to be done on this, and only one or two remarks may here be added.

(3.4.1) Quantities proportional to metron-content

In the first place, it is impossible for a quantity proportional to metron-content to be represented as having the value zero. It can only be shown as occupying the first interval on the metron-scale. It would appear in particular that the concept of "zero energy" cannot find a representation in this form. It may even be doubted whether it is a scientifically legitimate concept, since it has never had and can apparently never have observational illustration. The smallest mean value which can be attached to such a concept would seem to be one half the magnitude of the first interval it can be alleged to occupy.

(3.4.2) Coordinate versus magnitude

Secondly, there may often be a fundamental distinction between representations which may depict the same physical quantity, according to whether it is an identifying coordinate or a measured magnitude. A good example is provided by the case of radar. Woodward's[10] treatment of radar information shows in effect that the proper-scale of time as a measured magnitude (i.e. delay of an echo), is linear in that magnitude. The precision with which it may be subdivided is proportional to the square root of what Woodward has called the numerical energy, (which is proportional to our metron-content). In other words the number of metrons is proportional to the square of the number of intervals on the time-scale showing the delay.

On the other hand the flow of metrons enables us conceptually to subdivide the scale of time as a coordinate into a number of intervals which is proportional to metron content. If a single pulse of a given duration yields a certain number of metrons, we may picture these as arriving at small intervals of time which are inversely proportional to the total number in that given time. Such an interval may in general be called a conceptual unit of time (or

Conceptual units

* This figure differs by a small factor from that given in P.M., which applied to a proper-scale graduated in steps equal to the standard deviation.

whatever coordinate is involved), and represents the interval of time which can be conceptually associated with each metron, on the time-coordinate scale.

In other words, it is a pure accident, so to speak, that the result of a series of observations is a sequence of time - measurements distributed in time and therefore identified also by time-coordinates.

(3.4.3) Theoretical representations

Thirdly, when making theoretical abstract representations of magnitudes, we may have nothing corresponding to an observed standard deviation to provide us with a natural unit of magnitude. Yet if our representation is to be made at all we must be able somehow to define a scale-unit. A preliminary investigation suggests that in every case in which the concept of magnitude arises in theoretical physics, it does so in a context in which at least two magnitudes of the same kind appear.

Two examples may be cited: (a) if one's interpretation of Eddington's calculation of the radius of the universe is correct, he appears in our language to take the uncertainty of the centroid of N particles as the scale-unit in terms of which to specify the radius of the universe containing them on grounds very similar to those advanced here. (b) It is interesting to observe that if we formulate the concept "The mutual energy of two charges $\pm$ e", the formulation specifies two lengths: the separation r between the charges, and the wavelength λ of the radiation to which the energy is equivalent. The smallest interval conceptually defined by "a wavelength λ" in terms of coincidence-relations, is[x] the 'radian-length' $\lambda/2\pi$. It may be easily verified that irrespective of the energy, the ratio of these two lengths, $\lambda/2\pi r$, is a constant $hc/2\pi e^2$. This is the famous fine-structure constant, which is known to be very nearly the integer 137, and was asserted by Eddington to be exactly that integer. Its role in terms of our information pattern is at least suggestive.

(3.5) The information-vector

(3.5.1) Information-space

We have seen that the metron-content of a combination of two measurements is the sum of their separate metron-contents. We have seen that in the case of many magnitudes of interest (particularly vector quantities such as voltage, current, velocity and the like) which are linearly represented on a proper-scale, the metron-content is proportional to the square of the magnitude concerned.

Information space

This suggests a convenient way of representing the information-content of a series of observations of such a magnitude. Each successive independent reading or logon will have a certain metron-content and will be represented by a certain proper-scale reading equal to the square root of this. Let us now take a (mathematical) space, which we can call the information-space, with as many dimensions as we have logons, and take the proper-scale reading of each logon as one coordinate of an information-point or volume-element in the space. Alternatively we may think of it as one vector-component of an information-vector linking the information-point to the origin. With this representation, the total metron-content of our result

---

[x] P.M. para. 8 (a).

is the square of the length of the information-vector, and is the sum of the squares of its components, as required.

Specifying this vector thus specifies the complete result, since its dimensionality indicates the logon-content, its length the square-root of the metron-content, and its orientation the relative magnitudes of the individual logons or the form of the result. The angle between two information vectors - or its cosine - quite literally measures the bearing of one or the other.

If now we apply some method of linear analysis to our result, so as to represent it as the sum of a number of components, each of these components will be defined in form by a corresponding vector-function or ray in the information-space. The metron-content of that component is then simply the square of the projection of the information-vector on the corresponding ray. A complete analysis into a fresh set of orthogonal components would be represented by projection of the vector on to a new set of orthogonal rays. This would amount to a rotation of the axes, and the sum of the metron-contents of the new components would be the same as the original total metron-content. This accords with our axiom that complete reformulation leaves the total number of logical elements unaltered.

(3.5.2) 'Barter' of metrical and structural information

It should perhaps be said that this vector-form of representation is not of universal application; but it provides a useful and illuminating picture of many common processes in experimentation. The process of narrowing bandwidth, for example, amounts to rotating the information-vector so that it lies in or near a subspace of fewer dimensions. The average metron-content per logon is thus increased, in quantitiative accord with the well-known relation between bandwidth and signal: noise ratio. The telephone engineers' practice of using 'top-lift' filters in order to increase bandwidth, is represented by the converse process and has the corresponding effect of reducing precision. The essential point which is brought out is that no artifice of manipulation can yield a result having a greater metron-content than the original, while if the transformations are not thermodynamically reversible, the total metron-content may be reduced. If they are reversible however, we can infer from the equivalence of metron-content and physical entropy that metrical information cannot be destroyed by a reversible transformation. An example of such a transformation is the effect of a filter comprising only purely reactive components.

It is not always obvious that desired rotations of the information vector can be realised; but the picture in any case sets a useful upper limit to what we may legitimately seek to attain. Logon-capacity (resolving power) can in principle be increased at the expense of average metron-content, but the upper limit is set by the length of the information vector. When its dimensionality has

been increased to the point at which only one metron is available per logon, (in practice well before then), the corresponding statement has reached its lower limit of probability and its maximum of useful complexity.

The reader will realise that to speak of 'exchanging' metrical for structural information does not imply that they are at all of the same kind, but merely that the one 'kind' of information is shared out among units of the other.

(4) SELECTIVE INFORMATION

(4.1) The choice specified by an experiment.

We must now briefly see how the concepts of 'information' we have so far considered are related to the third use of the term as measuring complexity of choice. We adopt here a new standpoint, which we briefly considered in paragraph 1.4.1, and view an experiment not primarily as giving us a description, but as instructing us to make a selection. What we select ceases to concern us; we ask now rather "from how many competitive possibilities has this experiment selected a result?"

Here our information-space is often useful. We recall that its dimensions were calibrated as proper-scales, so that any of the quantised intervals were physically equally likely to be occupied. In any given circumstances, therefore, we can calculate the total number of possible equiprobable positions of the information-point, and think of our experiment as having selected one out of this number of possibilities - namely the actual result observed.

Selective Information - content

Binary digit: Bit

Now the most economical way in which a selection can be made out of M possibilities is in a series of 'yes-or-no' choices between equally large groups of possibilities at each stage - the now familiar process of binary selection. The number of such binary choices, N, is the nearest (higher) integer to $\log_2 M$. If then our information space contains M possible information-points or cells, the selective information - content of a result is defined to be the number N of binary choices determined when it is observed. These have been called binary digits or 'bits' of information - i.e. of selective information. The calculation of this number under various conditions has been performed elsewhere[*], and need not concern us here. Evidently, however, we shall from our new standpoint regard an experiment as most meritorious when designed to make the number N as large as possible. Increasing either the length or the dimensionality of the information vector will increase it, but the second is clearly more profitable unless we already have only one metron per logon. Once again, however, we see that a barter-principle applies. We can increase the selective information-content of a result either by increasing its dimensionality (e.g. by increase of bandwidth) or its total metron-content (e.g. by increase of total energy). This is what writers have in mind when they say that 'energy can be traded for bandwidth'.

---

[*] See P.M. para. 9 and ref. 6.

(4.2) Choosing between different measures of information

The above criterion is in one sense more fundamental than either of the previous complementary pair. We have seen that it is the chief one of interest to the communication engineer, because he is prepared to design a code-system by which each information point identifies for him in his ensemble a representation as complex as he wishes. It is also a good general-purpose criterion of merit enabling us to compare two results differing widely in form. Its use is, however, conditional on our agreeing that what matters, roughly speaking, is the volume of information-space rather than its shape;

Unforeseeableness

the 'unforeseeableness' of a result rather than its precision or degree of resolution. If we are blessed with ingenuity and cooperation on the part of nature, this may be the dominant consideration. "Give me the bits, and resolving-power will take care of itself" might perhaps be suggested as a challenging watchword, already translated into action by communication engineers.

On the other hand, when a physicist narrows bandwidth or performs some other equivalent averaging operation to reduce the logon-capacity of his apparatus, it is an irrelevancy to point out to him that he is reducing the selective information-content of his result. He is interested in something quite different - namely the precision of the measurement he is making. He knows in fact from consideration of the information-pattern (and knew intuitively all along) that any intelligent steps to sacrifice logon-content can be rewarded with increased metron-content, other things being equal. Reproducibility, not amount of detail, is his relevant criterion.

When conversely a physiologist reduces the inertia of a galvanometer to enable it to respond to higher-frequency components of electroencephalographic waveforms, it is primarily the number of independent ordinates per second that interests him. The 'selective' criterion, though less inappropriate this time, would again not represent what he has in mind, since it takes into account also the precision of his ordinates.

It is thus clear that all three measures are appropriate for different purposes. There is certainly no question of advocating the first two as against the third, or vice versa. Our choice should be as flexible and intelligent as it is between volume, area or length as a measure of 'size' - and if possible, as dispassionate.

(4.3) The distinction between 'unforeseeableness' & randomness

The 'unforeseeableness' which we have taken as one index of merit in a result is not to be confused with statistical randomness in the result. The distinction is much obscured in the literature, and the impression could be gained that it was a virtue in a 'noisy' result, that one could never predict what it would do next. Mathematically it is true that a 'white-noise' signal is the least predictable in a given bandwidth; and mathematics

unilluminated by physical insight appears to attribute more 'information' to any waveform the nearer it approximates to white noise in structure.

The fallacy of course lies in failure to ask about what one is uncertain. The 'unforeseeableness' mentioned in the definition of selective information refers to members of the particular array of possibilities (and representations thereof) which the observer is assumed to have foreseen and prefabricated. The greater their number, the harder it is to foresee which will turn up, and the greater the merit of a result which singles out one.

But the larger the random noise component in a result, the smaller will be the number of significantly distinct representations which the observer will have prepared. His aim is not to describe exactly what he observes, but what he can assert with reasonable probability to be the case - i.e. to be reproducible. He is no doubt perpetually being surprised by the noise-structure of his result. But he gains no selective information from it because he has prepared nothing in his ensemble which it can tell him to select. It is only when in a communication-system someone deliberately reproduces a recorded (and recognisable) noise-signal and transmits it as a code-signal, that its noise-like properties have any merit.

(4.4) Distinction between signals and non-signals

Finally it must be remembered that receiving a communication signal differs fundamentally from observing a physical sequence which is not a signal, in that the former may be known to favour certain regions of information-space which on physical grounds have no abnormal probability-density. It then becomes necessary either to warp the information-space so as to restore its "equiprobability" before calculating its volume for the purpose above, or to apply the more detailed treatments of statistical mechanics, as is in fact done in communication theory. (See Glossary, and/or refs. 3, 5, 6, 7).

(5) IMPLICATIONS

(5.1) Practical aspects

The chief practical merit of this approach is in its provision of quantitative criteria of different kinds for assessing how far an experiment has come short of the ultimate limit. With respect to coordinate-resolving-power (logon-capacity) we have seen that as long as the conceptual unit of coordinate associated with one metron (3.4.2) is smaller than the structural scale-unit which forms the 'base' of a logon, it may be legitimate to apply ingenuity towards increasing logon-content at the expense of metron-content.

Conversely an increase in metrical precision (metron-content) can in principle always be sought if the dimensionality of the information-vector is unnecessarily high; but no artifice of manipulation can wring more precision out of a result than is given by the total metron-content. This is governed fundamentally by the number of conceptual units in the tract of coordinate-space employed. In general, as we have seen, each metron is associated with a definable conceptual unit of space or time or space-time, so that the absolute limit to metron-content is set by the number of

these units allocated to the experiment. (For example in the case of power-measurement considered in paragraph 3.3.5 the conceptual unit of time is 0.91kT/W. An analogous unit of area can be defined for a space-statistic such as the density of a photographic plate).

Our first question then is always of the firm: "Have I here surplus metrical (or structural) information? If so is there not some way of increasing the one by sacrificing the other?" We have thus a quantitative limit to 'legitimate aspiration', which is safest when used negatively, but may serve positively as a useful stimulus.

The ability to calculate metron-content is also useful when considering the statistical matching of one part of an experiment to another, a problem whose solution is familiar to statisticians. For instance, if a 'weak link' is known to yield only a certain metron-content i_o, it is possible to estimate the time and/or space (or energy) which it is worthwhile to devote to each of the remaining links, and to gain in overall metron-capacity (metron-content per unit of space-time) by deliberately designing these so as to barter accuracy for speed or compactness. It is wrong to say (e.g. that it "does no harm" to use a galvanometer which is unnecessarily sensitive; for a less sensitive one could have a more rapid response, and might allow several measurements each yielding i_o to take place in the same time.

Metron-capacity

Other illustrations of this approach have been given in P.M., and an elegant extension of it is presented in Dr. Gabor's paper in this symposium.[3]

At the most general level our attitude has been to regard an experiment as specifying one out of a limited number of possibilities. The merit of an experimental method may then be judged in general by the speed or ease with which it enables us to identify the information-point specified.

(5.2) Theoretical aspects

It is easy to overrate the significance of any theoretical conclusions reached here. We have, however, seen that wherever a compromise has to be struck in experimentation, it becomes possible to identify a definite quantum associated with the measurement concerned. The example chosen was that of entropy (para. 3.3.5), but the general principle can be stated thus: each unit of metrical or structural information is associated with a definite tract of the corresponding co-ordinate-space (time, area, etc.), by a relation of the form: $\Delta y. \Delta q > K$

Here Δy and K have quite different meanings according to whether Δq is a structural scale-unit or a conceptual unit defined by the density of metrons on the q-axis.

(a) In the first case y is the Fourier - transform of q (as frequency is of time) and K is a number calculable a priori. Δy is then the bandwidth of our apparatus or its analogue.

(b) In the second, y is that function of the measured quantity which is proportional to metron-content and Δy is the interval on the scale of y corresponding to one metron. K then depends on the nature of the environment.

The examples we have met (in para. 3.3.5) were

(a) $\Delta f . \Delta t \gtrsim 1/2$

(b) $\Delta W . \Delta t \gtrsim 0.91kT$

The present point is that these relations from our viewpoint appear to be shorn of any element of frustration. For we cannot, in this logical language, speak of what we cannot observe. In terms of atomic propositions, it appears that the dilemma which one normally associates with uncertainty-relations cannot be formulated. They may be thought of as expressing an informationaxiom.

We have seen that the principle of the indestructibility of elementary proporitions may not be unrelated to fundamental theory; but this aspect of the subject awaits further investigation. It is, however, certain that any two representations which are reversibly equivalent must have the same number of elements, and this general conclusion cannot be without its implications in any field where the same concept may be defined in two independent ways.

(5.3) The mechanism of thought

Only very general relevance can be claimed for this approach, to the problem of human reasoning and the possibility of explaining or imitating it mechanically. It seems however to be highly suggestive in a number of ways, apart from its more technical applications to calculation of the information-capacity of neural elements.

The complementary roles of metrical and structural information in particular, as suggested in P.M., seem to indicate a way in which a machine embodying both 'analogue' and digital' principles could be made to operate in a manner more closely analogous to that of the normal human mind than the behaviour of a digital computer appears to be. Briefly, it is suggested that a mechanism in which the degree of confidence in propositions, - the probability of excitation to belief - was continuously variable, could hold and act on information of the uncertain statistical kind represented by our information-vector, in a way analogous to our own reactions to the same kind of information. Normal human thinking then would be regarded not as a strictly logical process, but as a quantal approximation to it, - a process of only limited determinacy governed by transition-probabilities corresponding to metrical information. Such a mechanism would essentially comprise two interlocked networks, one, roughly speaking, handling propositions, the other handling the probabilities to be associated with them.

Further discussion of this point, however, will be more appropriate at a later stage in the symposium.

(5.4) Conclusion

We must now seek to draw together some of these diverse threads. What has been said here contains little that is now new. It presents practically no new results, though it provides variants to the usual derivations of some. Its aim has been chiefly to suggest a way of looking at experimentation, and a generalised language, which may give a more ready or more stimulating insight into the principles underlying experimental physics, and which appear to provide specially simple and fundamental grounds for expecting certain well-known relations to take the form they do.

If one were to sum up our present thesis in a sentence, it would be this: A given experimental situation is equivalent to the specification of a certain number of elementary propositions; I may judge the extent to which I have exhausted its possibilities by the fraction of that number of propositions which appear in my statement of the result.

Judged by this criterion alone it is apparent that most techniques fail lamentably; but it may often prove illuminating to pursue the reasons for the apparent failure; the effect may be in each case to disclose a 'barter principle' which if not unsuspected was hitherto only intuitively known.

The advantage offered by the two or three new terms introduced lies perhaps mainly in their generality. It is possible now to formulate general principles which apply alike to the design of optical instruments, galvanometers, and communication-channels, for example, and to see the origin and validity of many analogies in identity of basic logical form.

It must be confessed by way of warning however that in most cases one's conclusion is a rather shamefaced admission that "I ought to have seen this before". Perhaps this is as it should be, in a primarily linguistic contribution to the field of experimentation.

(6.) ACKNOWLEDGMENTS

The present paper is a revised presentation of the ideas contained in the paper "P.M." and represents no substantial change except in the terminology connected with metrical information, where the opportunity has also been taken to clarify the distinction between what is now termed the proper scale, and the conceptual scale of metron-content. The acknowledgments in P.M. are therefore still due, while the author's indebtedness to the recent writings of others in the field will be obvious.

Sincere thanks are also due to all those who have raised difficulties in the course of past discussions, - especially (with apologies) to any who may find their spontaneous objections presented herein as the utterances of the advocatus diaboli.

REFERENCES:-

1. Fisher, R.A. - "The design of Experiments", p. 188 London: Oliver & Boyd, 1935.

2. Gabor D. - J. Inst. Elec. Engrs. 93 (III) 429, 1946.

3. Ibid - "Communication Theory & Physics", Symposium paper.

4. MacKay, D.M. - Phil. Mag. Ser. 7, 41, 289, 1950. (Referred to as P.M. herein)

5. Shannon, C.E. - B.S.T.J. 27, 379, 623, 1948.

6. Ibid - Proc. I.R.E. 37, 10, 1949.

7. Tuller, W.G. - Proc. I.R.E. 37, 468, 1949.

8. Wiener, N. - "Cybernetics", New York: John Wiley & Sons, 1948.

9. Wittgenstein, L. - "Tractatus Logico-Philosophicus" Kegan Paul, 1922.

10. Woodward, P.M. - "Theory of Radar Information", Symposium paper.

THE STATISTICAL APPROACH TO THE ANALYSIS OF TIME-SERIES.

by
M.S. Bartlett.

Summary

The problems of statistics are broadly classified into problems of specification and problems of inference, and a brief recapitulation is given of some standard methods in statistics, based on the use of the probability p (S/H) of the data S on the specification H (or on the use of the equivalent likelihood function). The general problems of specification and inference for time-series are then also briefly surveyed. To conclude Part I, the relation is examined between the information (entropy) concept used in communication theory, associated with specification, and Fisher's information concept used in statistics, associated with inference.

In Part II some detailed methods of analysis are described with special reference to stationary time-series. The first method is concerned with the analysis of probability chains (in which the variable X can assume only a finite number of values or 'states', and the time t is discrete). The next section deals with autoregressive and auto-correlation analysis, for series defined either for discrete or continuous time, including proper allowance for sampling fluctuations; in particular, least-squares estimation of unknown coefficients in linear autogressive representations, and Quenouille's goodness of fit test for the correlogram, are illustrated.

Harmonic or periodogram analysis is theoretically equivalent to autocorrelation analysis, but in the case of time-series with continuous spectra is valueless in practice without some smoothing device, owing to the peculiar distributional properties of the observed periodogram; one such arithmetical device is described in § 7. Finally the precise use of the likelihood function (when available) is illustrated by reference to two different theoretical series giving rise to the same autocorrelation function.

(I) GENERAL PRINCIPLES.

(1) STATISTICAL SPECIFICATION AND STATISTICAL INFERENCE.

R.A. Fisher (1925) has listed the problems of statistics under three headings:-

(a) the problem of specification of the theoretical probability model which is to represent some actual statistical phenomena, apart perhaps from some unknown parameters;

(b) the problem of estimation of the unknown parameters from statistical data;

(c) the problem of distribution of these estimates, or of any further statistics calculated for purposes of inference.

In some ways, however, it is more convenient to group these problems a little more broadly, under just the two headings:-
firstly, the problem of specification;
secondly, the problem of statistical inference (which includes both (b) and (c) above).

These two broad classes of problem are now largely, but not altogether, separate problems. The first requires, in addition to a knowledge of possible mathematical models, a practical knowledge of the physical phenomena to be represented. It is not of course completely independent of the second class of problem, for one of the latter's functions will be

to check the adequacy of the specification. The more detailed the specification the narrower is the inference problem, but at the same time such a detailed specification may prove untenable as a representation of the date.

Where possible, the specification or probability model [X] H should specify precisely the probability of the data S, (and of any alternative set of data S' which might have arisen under the same conditions). I shall denote this probability by p (S/H). If S refers to observations which have a continuous range of variation, p will be strictly zero, but we may consider alternatively the density function f (S/H). The function p or f is called the likelihood function and I shall denote its (natural) logarithm by L. H may only be known to be in some class; in particular each possible H may correspond to a particular set of values θ^i of a number of unknown parameters or constants.

It will be useful now to summarize briefly some important formulae in the theory of statistical inference; further details may be found in Cramer (1946) or M.G. Kendall (1946a). Fisher's information function I (θ) is defined by

$$I(\theta) \equiv E\left\{\left(\frac{\partial L}{\partial \theta}\right)^2\right\} \qquad (1)$$

where one parameter θ (the possible values of which are assumed to have a continuous range) is to be estimated, and E denotes expectation with respect to the probability distribution p (S/H). In the case of more than one unknown, we have the information matrix

$$I^{ij} \equiv E\left\{\frac{\partial L}{\partial \theta^i}\,\frac{\partial L}{\partial \theta^j}\right\} \qquad (2)$$

Under suitable conditions (the important practical condition to check is that the operators E and $\partial/\partial\theta$ commute, this usually implying that the possible range of the observations S is independent of θ) we have for any estimate T(S) of θ

$$E\left\{(T-\theta)^2\right\} \geq (1+\partial b/\partial\theta)^2/I(\theta) \qquad (3)$$

where $b \equiv E(T)-\theta$ is the bias of T. For unbiased estimates b is zero, and (3) reduces to

$$\sigma^2 \geq 1/I(\theta), \qquad (4)$$

where σ^2 is the variance of T. The condition for the equality sign in (4) is that

$$\partial L/\partial\theta = I(\theta)\,[T-\theta]\,. \qquad (5)$$

In the case of more than one unknown I shall restrict myself for simplicity to unbiased estimates. We then have the result that the variance-covariance matrix of the set of estimates T^i of θ^i is 'bounded' below by the inverse of the information matrix, by which is meant in particular that σ_i^2 is not less than the corresponding diagonal element of the matrix, (and more generally that similar inequalities hold for any linear transformations of the estimates). The set of conditions for equality in these relations is

$$\frac{\partial L}{\partial \theta^i} = I^{ij}(T^j-\theta^j), \quad \text{(summation convention)}. \qquad (6)$$

X It is sometimes important to distinguish between H and the 'structural' model leading to H. If two different structural models give rise to the same H, no statistical analysis can of course discriminate between them.

The 'maximum likelihood' estimates ϱ^i are defined as any set of values of ϱ^i that maximize L for given S. In particular when $\partial L/\partial \varrho^i$ exist, they satisfy

$$\frac{\partial L}{\partial \varrho^i} = 0. \tag{7}$$

Under some further conditions (these have usually included the assumption that S consists of n independent observations) the estimates ϱ^i have the property that as the number n of observations is increased they tend to be normally distributed about ϱ^i with their variance-covariance matrix the optimum compatible with the above results. Thus while uptimum estimates in the above variance sense may not exactly exist (if they do, they are identical with the maximum likelihood estimates) equation (7) will provide estimates which have these optimum properties at least asymptotically.

The relation

$$E\left\{\frac{\partial L}{\partial \varrho^i}\,\frac{\partial L}{\partial \varrho^j}\right\} = E\left\{-\frac{\partial^2 L}{\partial \varrho^i\,\partial \varrho^j}\right\} \tag{8}$$

usually holds, and is then often convenient for evaluating I^{ij}. The asymptotic approximation

$$-\frac{\partial^2 L}{\partial \varrho^i\,\partial \varrho^j} \sim I^{ij} \tag{9}$$

may also be useful.

In some cases where $p(S/H)$ cannot be completely specified estimates can be chosen with reasonable properties for a wider class of H. A well-known example is that of 'least-squares' estimates, which are, for a particular class of H, unbiased with minimum variance in the group of estimates obtained by taking linear combinations of the observations.

The above formulae refer to statistical estimation. In the case of statistical tests it is known (as is intuitively obvious) that the best criterion for testing one hypothesis or model H_o against a rival H is the 'likelihood ratio' p/p_o (or f/f_o if only densities exist), where for brevity I write $p_o \equiv p(S/H_o)$, etc. Instead of p/p_o we may equivalently consider $L-L_o$. If moreover p has the form

$$p \equiv p(S/H) = p(U/H)p(S/U),$$

where the last factor does not depend on H (or the aspects of it which are in doubt) and U denotes a reduced set of statistics obtained from S, then p/p_o is a function only of U, which are said to be a set of sufficient statistics in regard to H. In particular U may be a single statistic T.

The method of using p/p_o will in general depend of the situation, especially if the hypothesis H rival to H_o is merely one in a class. But in the probability space of S a particular region will be favourable to H_o against H, and this is defined by some condition $p/p_o < \lambda$. When a sufficient statistic exists, the region will be defined in terms of critical values of T. It should be noted that, if equation (5) holds, T is sufficient (but the converse need not hold). In some cases when H and H_o are not completely specified, but depend on further 'nuisance parameters' which are unknown, it is possible and advisable to remove these completely by considering the conditional probabilities $P(S/H,W)$, where W is a set of sufficient statistics for the nuisance parameters. When these cannot be removed exactly, they may be removable approximately by the substitution of their maximum likelihood estimates.

Suppose now we wish to test the 'goodness of fit' of a model specified entirely (apart perhaps from further assumptions which are maintained throughout) by the set of parameters θ^i, ϕ^j, where θ^i ($i = 1...r$) are believed known, but ϕ^j ($j = 1...s$) unknown. The alternative class of hypotheses for comparison is both θ^i, $\phi^j \equiv \psi^m$, say, unknown. For a useful asymptotically valid test, we may substitute for ϕ^j in the former case their maximum likelihood estimates $\hat{\phi}^j_\theta$ (θ^i being given) and for both θ^i, ϕ^j their simultaneous estimates in the latter case. For $\hat{\psi}^m - \psi^m$ small,

$$L(\psi^m) \sim L(\hat{\psi}^m) - \tfrac{1}{2}(\hat{\psi}^m - \psi^m)\left(-\frac{\partial^2 L}{\partial \psi^m \partial \psi^n}\right)(\hat{\psi}^n - \psi^n)$$

$$\equiv L(\theta, \phi) \sim L(\theta, \hat{\phi}_\theta) - \tfrac{1}{2}(\hat{\phi}_\theta - \phi)\left(-\frac{\partial^2 L}{\partial \phi^j \partial \phi^k}\right)(\phi^k_\theta - \phi^k),$$

whence by subtraction

$$-2\,L(\hat{\psi}^m) - L(\theta^i, \hat{\phi}^j_\theta) \sim \chi_1^2 - \chi_2^2 \sim \chi_3^2 \quad , \tag{10}$$

where, if the standard asymptotic properties for the maximum likelihood estimates hold, χ_1^2, χ_2^2 and χ_3^2 are approximately χ^2 quantities (sums of squares of independent normal variables with zero means and unit standard deviations) with degrees of freedom (number of variables) r+s, s and r respectively.

(2) STATISTICAL SPECIFICATION OF TIME-SERIES.

If an actual time-series has a random, statistical or stochastic element in its make-up, it becomes what is called a stochastic process. Any realisation $x(t)$, owing to the random element, will in general differ from another; regarded as a random quantity it will be written $X(t)$. The complete probability specification for $X(t)$ must include a knowledge of the simultaneous probability distribution of $X(t_1), X(t_2), \ldots$ at any times $t_1, t_2 \ldots$; it is often possible to build up such a distribution to any required order from a knowledge of the mechanism of the process. For some purposes it is more convenient to adopt the alternative representation $X(t) = f(t, \Omega)$, $x(t) = f(t, \omega)$, to denote that $X(t)$ is a function of t depending on a random quantity Ω. As Ω has to represent all the randomness in $X(t)$ from $t = -\infty$ to ∞, it is in general a random quantity of a rather complicated and abstract kind. However, I do not need to consider the mathematical rigourisation of such ideas, carried out by Kolmogoroff, Doob, Wiener and others.

If X can only take a finite or enumerable number of values, corresponding to a similar number of 'states', I shall call $X(t)$ a probability chain, whether or not t takes a sequence of values or a continuous range. If the distributional properties of $X(t)$ do not depend on the absolute value of the time, then $X(t)$ is called stationary*. Many physical processes are of this kind, and I shall mainly consider time-series of this type, especially in the later exposition of some methods of analysis. The treatment of non-stationary time-series does not introduce any new principle, but it is of course in general more complicated.

In dealing with stationary time-series, especially for X taking a continuous range of values, it is often sufficient to concentrate on its linear correlation properties. Correspondingly, in the correlation (or harmonic) theory of these processes, it is sufficient to assume stationary 'to the second order', that is, in addition to the mathematical expectation or stochastic average $E\{X(t)\} = m$ (constant), which for convenience is put zero, it is assumed that the autocovariance function

$$E\left\{X^*(t)X(t+\tau)\right\} = w(\tau) \quad . \tag{11}$$

* For further details see, for example, Levy (1948), especially Ch. IV. The present summary is partly based on the introduction to my paper (Bartlett, 1950), where references to original sources (Khintchine, Wold, Wiener, Cramer and others) will be found.

depends only on the interval τ ($X^{*}(t)$ denotes the complex conjugate of $X(t)$, though in applications $X(t)$ is usually real). When X is standardized by dividing by its constand standard deviation σ, assumed finite, equation (11) gives the autocorrelation function $\rho(\tau) \equiv w(\tau)/\sigma^2$.

For stationary time-series defined for integral values of t any theoretically valid autocorrelation function $\rho(\tau)$ has the form

$$\rho(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega} \, dF(\omega) , \qquad (12)$$

where $F(\omega)$ has the form of a cumulative distribution function defined between $-\pi$ and π), and is the integrated spectrum of the process. For t taking a continuous range, it will be assumed that $X(t)$ is continuous in mean square, that is,

$$\underset{h \to 0}{\text{Lt}} \; E\left\{ \left| X(t+h)-X(t) \right|^2 \right\} = 0 ,$$

a condition easily seen to be equivalent to $\rho(\tau)$ being continuous at $\tau = 0$. The corresponding theorem to (12) is then

$$\rho(\tau) = \int_{-\infty}^{\infty} e^{i\tau\omega} \, dF(\omega) \qquad (13)$$

where $F(\omega)$ is now defined from $-\infty$ to ∞ (for real processes, $dF(\omega)$ is symmetrical about 0, and only positive values of ω, representing real 'frequencies', need be considered, but throughout this paper I shall for convenience keep to the definition (13)). Any distribution function $F(\omega)$ must in general be a linear sum of three components; the first is a step-function, corresponding to a discrete spectrum, the second is absolutely continuous, and corresponds to a continuous spectrum, the third 'singular' component has no practical importance and will be assumed absent.

A series of independent disturbances (uniform 'noise') is well-known to give rise to a uniform spectrum. Strictly speaking, the above representation then breaks down in the case of time continuous, for $\rho(\tau) = 0$ except at $\tau = 0$; but this case may be handled by some convenient limiting procedure.

Corresponding to (12) and (3) there is an equivalent relation for $X(t)$ itself, associated with its harmonic analysis, namely

$$X(t) = \int_{-\pi}^{\pi} e^{it\omega} \, dZ(\omega) \qquad (14)$$

for integral values of t, and

$$X(t) = \int_{-\infty}^{\infty} e^{it\omega} \, dZ(\omega) \qquad (15)$$

In these formulae $Z(\omega)$ is an uncorrelated (orthogonal) process, and has the property

$$\Delta F(\omega) = E\left\{ Z^{*}(\omega+\Delta\omega) \, \Delta Z(\omega) \right\} = E\left\{ \Delta Z^{*}(\omega) \, \Delta Z(\omega) \right\} ,$$

where Δ denotes a 'first difference', and the stochastic Stieltjes integrals (14) and (15) are defined as mean square limits of appropriate stochastic sums.

Various other classifications are of importance. $X(t)$, even if stationary, may or may not be ergodic, and only in the former case can we infer its properties from a single realised series $x(t)$; the property of ergodicity will thus be assumed in relation to any aspect of $X(t)$ we may wish to investigate. $X(t)$ may be essentially indeterministic or contain a deterministic (permanently and exactly predictable) component. These aspects have been discussed by Wiener (1949) in connection with his theory of prediction, though, it should be noted, mainly in relation to linear correlation properties and linear prediction formulae. A useful sufficient condition for a process to be deterministic is that its autocorrelation function should be everywhere analytic.

A normal process is one for which the simultaneous distribution of $X(t_1)$ $X(t_2)$, ... at any number of points $t_1, t_2, \ldots$ is the multivariate normal distribution. Such a process is specified completely by its mean and autocovariance function. For stationary normal processes the $Z(\omega)$ in (14) or (15) above become additive processes (successive increments independent).

Further classifications and definitions, with particular reference to physical problems, have been summarized by Moyal (1949).

(3) STATISTICAL INFERENCE FOR TIME-SERIES.

In contrast with the theoretical specification of time-series, the theory of inference for time-series has not been considered very systematically, and only recently in relation to the theory of stochastic processes. Wiener's prediction and filtering technique is in one sense an exception, but the specification of the time-series is usually assumed already known, or inferred by ergodic principles from an infinitely long series. While there is no doubt that the general mathematical theory of stochastic processes, and in particular the theory of stationary processes, throws a powerful light on the possibilities of analysing time-series, it also raises many new problems connected with sampling fluctuations in the series. For some physical processes the 'length' of a series available for study may be as much as desired, whereas in other fields (especially in economics) the lengths of series available are severely limited. In all cases, however, the magnitude of sampling errors must be considered: this is most strikingly illustrated in periodogram analysis, where uncritical estimation of spectral functions has led to sampling fluctuations that do not diminish with the length of series taken.

In Part II I shall discuss some particular methods of analysis of time-series. The greater complexity and newness of this field means that in spite of some exact results the sampling theory of time-series is still in a comparatively primitive state, and I shall be content on the whole to discuss asymptotic or "large-sample" methods. I shall, however, consider briefly in this section the general problem of statistical inference for time-series, making use of a recent fundamental paper by Grenander (1950), and referring mainly to those points in the theory that seem to me to be of importance in applications. Some particular estimation problems discussed by Grenander will also be re-examined at the end of Part II.

For time-series defined for integral values of t no difficulty of principle arises, as any length of series available constitutes a finite number of observations of the type already envisaged. The known theorems on the asymptotic properties of maximum likelihood estimates do not in general apply to dependent observations, and have to be extended. One such extension will be mentioned in the section dealing with probability chains. Another more practical difficulty is that the distribution and dependence of the observations may in many situations be imperfectly known or deliberately simplified, so that the precise formulation of the likelihood function is hardly feasible or useful. Methods then have to be employed which are of rather wide validity (like the use of 'least-squares' estimates, considered in the discussion of autoregressive time-series). However, in some physical processes at least this difficulty does not arise.

A further theoretical problem arises when we consider time-series defined for continuous time, and for which continuous time-records are available. It is then necessary to set up for such a record something that will stand for the likelihood function. This may be done if we can describe the record fully by a denumerable sequence of coordinates. It will be sufficient to consider two practically useful ways of doing this.

(i) For processes continuous in mean square we may consider the values at $\underline{n}$ points $t_1, \ldots t_n$, and then let $\underline{n}$ increase such that max $(t_r - t_{r-1})$ decreases to zero.

(ii) For processes for which $dX(t)$ is zero except at a denumerable number of random times $T_1, \ldots T_N$, $\underline{N}$ being also random (with probability one of being finite), we may specify the probability in terms of these times, the corresponding values of $dX(t)$, together with $X(0)$ and $\underline{N}$.

More abstract representations have been considered by Grenander, who has shown that it is legitimate to evaluate the likelihood ratio for the data $\underline{S}$ either from (2) (when appropriate), or as a limit, when $\underline{n}$ increases, from (1). This likelihood ratio, if evaluated for a fixed hypothesis H_o in the denominator, we have seen can be used in estimation problems. The exact or 'small-sample' theory of estimation summarized in the first section then still applies. The problem of elucidating minimum conditions for which the asymptotic properties of the maximum likelihood estimates hold becomes of course more formidable, and will not be discussed further here, though by analogy with the discrete probability chain case discussed presently it might be expected that for stationary ergodic non-deterministic time-series for which the dependence drops off sufficiently rapidly these properties will still hold.

(4) THE RELATION BETWEEN TWO INFORMATION CONCEPTS.

We have seen that the problems of specification of, and inference from, time-series are fairly distinct. Theoretically we may discuss the first without the second. Though the converse is not true, nor is it often true that the practical study of time-series can be separated from inference problems. Correspondingly, the information concepts used in specification and inference have been distinct and not interchangeable (in spite of an implication to the contrary by Wiener in Cybernetics (p. 76)).

In a specification of the uncertainty represented by the possible eventualities $\underline{S}$ in the specified probability distribution $p(S/H_o)$, it has been convenient in communication theory (see Shannon (1948)) to define the information (entropy) concept

$$J_o^o \equiv -E_o(L_o), \tag{16}$$

where for unambiguity I use E_o to denote averaging with respect to the probability distribution H_o. No inference problem exists and H_o is supposed known and definite.

In statistical inference problems I have noted above the use of Fisher's information matrix function I^{ij}, which is related to the change of $\underline{L}$ with H. $\underline{H}$ was only supposed known when the parameters θ^i were known, and I_o^{ij}, say, is then defined as $E_o\{(-\partial^2 L/\partial\theta^i\,\partial\theta^j)_o\}$ when $\underline{H}$ is in fact H_o.

In order to stress both the difference and the relation between these two concepts, I shall define a third 'information function' which includes both of these*. This function is defined as

$$J^o \equiv -E_o(L). \tag{17}$$

* Cf. Good (1950), § 6.9 who, however, did not notice the difference between $E_o(L)$ and $E(L)$, and thus also implied that the entropy concept could be used for inference problems. He has agreed with me that his last remark (vii) in this section requires correction. There is of course something to be said for his suggestion of calling $-L$ itself

... $\underline{H}$ 'near' H_o,

$$L = L_o + (\theta^i - \theta_o^i)\left(\frac{\partial L}{\partial \theta^i}\right)_o - \tfrac{1}{2}(\theta^i - \theta_o^i)\left(-\frac{\partial^2 L}{\partial \theta^i \partial \theta^j}\right)_o (\theta^j - \theta_o^j) + \ldots$$

and

$$-E_o(L) = -E_o(L_o) + \tfrac{1}{2}(\theta^i - \theta_o^i) I_o^{ij} (\theta^j - \theta_o^j) + \ldots \qquad (18)$$

provided $E_o\left\{(\partial L/\partial \theta^i)_o\right\} = 0$, as is usually true. That is, J^o_{oij} is the first 'constant' term in the expansion of J^o in powers of $\theta^i - \theta^o$ and I_o^{ij} occurs (as a rule) in the next term. This relation between I_o^{ij} and J^o may be expressed equivalently by the formula

$$I_o^{ij} \equiv \left(\frac{\partial^2 J^o}{\partial \theta^i \partial \theta^j}\right)_o \qquad (19)$$

Some further incidental comments on these formulae are:-

(i) We have seen that inference problems can be handled in terms of the likelihood ratio, or equivalently in terms of $L - L_o$. Thus we might define alternatively

$$I_o^{ij} \equiv \left(\frac{\partial^2 \Delta J^o}{\partial \theta^i \partial \theta^j}\right)_o \qquad (20)$$

where $\Delta J^o \equiv -E_o(L - L_o)$.

This use of $L - L_o$ should not be confused with the presence of an arbitrary constant which may be added to L or L_o and hence to J^o or J^o_o, when $\underline{S}$ refers to continuous observational variables, (corresponding to a change of coordinates). For example, if this arbitrariness is removed from $J \equiv -E(L)$ by comparison with the standard distribution H_o, the 'relative' information (entropy) function becomes

$$-E(L) + E_o(L_o) = J - J^o_o ,$$

whereas the function ΔJ^o above is $J^o - J^o_o$.

(ii) For independent sets of observations, all the functions defined above, L, J^o_o, J^o, I_o^{ij}, etc., are additive.

(iii) For ergodic stationary and non-deterministic time-series, it will be true fairly generally that for a long realisation we shall have asymptotically (if $\underline{H}$ is H_o).

$$L \sim E_o(L) \equiv -J^o. \qquad (21)$$

This result includes the two asymptotic approximations notes respectively in communication theory and in statistical inference,

$$-L_o \sim J^o_o , \qquad \left(-\frac{\partial^2 L}{\partial \theta^i \partial \theta^j}\right)_o \sim I_o^{ij}.$$

(iv) For time-series the rates of information may correspondingly be defined as $\underset{T \to \infty}{\text{Lt}}\ J^o_o/T$ and $\underset{T \to \infty}{\text{Lt}}\ I_o^{ij}/T$, or, in a single formula corresponding to (21),

$$\underset{T \to \infty}{\text{Lt}}\ J^o/T. \qquad (22)$$

Formulae (21) and (22) may when convenient be combined.

(II) SOME METHODS OF ANALYSIS.

(5) STATISTICAL ANALYSIS OF PROBABILITY CHAINS*.

Let us consider first the problem of estimation and goodness of fit for a 'finitely-dependent' probability chain, by which will be meant a chain whose probability dependence does not extend more than a finite number of intervals (k, say), time being assumed discrete. The case k=0 is a random sequence, the case k=1 is known as a Markoff chain (the general case k finite can also be interpreted as a Markoff chain by a more complicated definition of states). As practical examples of probability chains may be instanced the sequences of observational counts by Svedberg and Westgren in their classical experiments on colloidal particles in suspension (see Chandrasekhar (1943)). Thus 25 consecutive counts in one sequence of Westgren's (representing the number of particles in a fixed small element of volume) were:

2 1 1 1 1 1 0 2 2 1 1 1 2 3 2 3 0 0 0 0 0 0 1 1 0.

In general let the sequence be denoted by

$$S \equiv X_1, X_2, \ldots X_{n-1}, X_n,$$

where the suffixes refer to the order of the observations. The variable X can take s values denoted conventionally by the states 1,2,... s (s will be assumed finite, though in practice this is not always strictly true e.g. for the above example). A sub-sequence $X_h, X_{h+1}, \ldots X_{h+k-1}, X_{h+k}$ can thus take s^{k+1} 'values' specified by the simultaneous values of the $k+1$ X's. Let the frequency of any such specified 'value' $(i,j,\ldots q,r)$ be $N_{ij\ldots qr}$. For brevity the value $(i,j\ldots q)$ of the subsequence $X_h,\ldots X_{h+k-1}$ will often be denoted by u, and correspondingly the frequency $N_{ij\ldots qr}$ by N_{ur}. I shall write $p_{ur} \equiv p(r/u)$. Then the assumption of k-dependence readily gives

$$L = \sum_{j=1}^{k} \log p(X_j/X_1, X_2, \ldots X_{j-1}) + \Sigma_{u,r} N_{ur} \log p_{ur}$$

or as n increases,

$$L \sim \Sigma_{u,r} N_{ur} \log p_{ur} \qquad (23)$$

Maximizing L with respect to p_{ur} and remembering that $\Sigma_r p_{ur} = 1$, we find for n large enough the estimates

$$\hat{p}_{ur} = N_{ur}/N_u, \quad (N_u = \Sigma_r N_{ur}). \qquad (24)$$

If we know the probabilities p_{ur} exactly, the goodness of fit criterion for checking this hypothesis (against the alternative class of any probability chain of maximum dependence k) becomes

$$-2(L-L_{max}) \sim -2\Sigma_{u,r} N_{ur} \log(p_{ur} N_u / N_{ur})$$

$$= 2\left[\Sigma_{u,r} N_{ur} \log(N_{ur}/m_{ur}) - \Sigma_u N_u \log(N_u/m_u)\right], \qquad (25)$$

where $m_{ur} = nP_{ur} = nP_u p_{ur}$, $m_u = nP_u$, P_{ur} and P_u denoting absolute probabilities of the 'values' (u,r) and u. In the independent case k=0, (23) becomes exactly $\Sigma_r N_r \log p_r$ and (25) $2\Sigma_r N_r \log(N_r/m_r)$. It is well known that the latter expression is asymptotically a χ^2 with s-1 degrees of freedom, and is equivalent, to the same order of approximation, to the more familiar expression $\Sigma_r (N_r - m_r)^2/m_r$. The degrees of freedom for (25) are $s^k(s-1)$, (s-1 for each set of estimates (24) for given u). To infer further that the expression in (25) for $k > 0$ has also an asymptotic χ^2 distribution, we require to know that the estimates (24) have the required asymptotic

---

* This section summarizes a paper not yet published (Bartlett, in press).

properties given in §1, although we are considering now dependent sequences (this requirement would be necessary even if the actual sequence is independent). This may also be shown under the further assumption that the probability chain is 'positively regular', by which is meant that the ultimate probability distribution of the different states exists* such that each state has a non-zero probability, (this assumption implies ergodicity and ultimate stationarity).

The efficiency of the estimates (24) implies automatically that they are consistent (tend to p_{ur} in probability), whence it follows that N_{ur}/n are consistent estimates of $p_u p_{ur}$, and

$$-L/n \longrightarrow - \Sigma_{u,r} p_u p_{ur} \log p_{ur} \qquad (26)$$

in probability, an important formula in communication theory (Shannon, 1948; cf. also §4). It may be deduced also from L that the information matrix for the p_{ur} is asymptotically equivalent to that for multinomial probabilities p_{ur} (u fixed) from $E(N_u)=m_u$ independent observations, with, moreover,

$$E\left\{ -\partial^2 L/\partial p_{ur}\, \partial p_{vq} \right\} = 0, \quad (u \neq v),$$

giving the variances and covariances of the estimates $\hat{p}_{ur}$ as standard multinomial formulae, with the further results $\mathrm{cov}(\hat{p}_{ur}, \hat{p}_{vq}) \sim 0$. The fluctuation formulae for the N_{ur} may also be deduced, either indirectly from the above results, or by other more direct methods.

If the probabilities p_{ur} are not known exactly, but only in terms of some m parameters α_ν, then the estimates of the α_ν are given asymptotically by the equations

$$\Sigma_{u,r} N_{ur}\, \partial \log p_{ur}/\partial \alpha_\nu = 0, \quad (\nu = 1, \ldots m) \qquad (27)$$

From the principles indicated in §1, this implies a loss of m degrees of freedom for the χ^2 goodness of fit criterion.

As a numerical illustration an artificial Markoff chain sequence of 2400 items was constructed from Tippett's random numbers according to the transition probability matrix

$$\begin{pmatrix} 0.625 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.375 \\ 0.125 & 0.25 & 0.375 \end{pmatrix}$$

(where each column represents the probability distribution arising from a given state). The observed and expected frequencies for N_{ur} (given k=1) were ✓ :

			Total
599 (599.50)	240 (222.67)	112 (137.03)	951 (959.20)
226 (239.80)	483 (445.34)	222 (205.54)	931 (890.69)
127 (119.90)	207 (222.67)	182 (205.54)	516 (548.11)
Total 952 (959.20)	930 (890.69)	516 (548.11)	2398 (2398)

* This distribution would not exist if the chain were deterministic i.e. 'cyclic' (see Frechet (1938)). For intermittent observations from a chain defined for continuous time, this case is automatically excluded.

✓ For this example I am indebted to Mr. B.J. Prendiville, who is making a detailed numerical examination of this test. (He is including a check of this model against the wider alternative k=2; this explains why the total frequency is only 2398). Mr. Prendiville has also been using this method to examine the adequacy of a Markoff chain fit to the colloidal particle counts.

The log formula gives

$$\chi^2 \text{ (6 degrees of freedom)} = 15.73 - 3.68 = 12.05,$$

and the quadratic χ^2 formula gives the approximately equal value

$$\chi^2 = 15.44 - 3.67 = 11.77$$

This value indicates a good enough fit, though as the 1 in 20 significant value of χ^2 is 12.59, the agreement is not quite as close as would be expected. It should also be noted that while the marginal frequencies are in accordance with their expectations if judged by the usual χ^2 quantity (3.68 with 2 degrees of freedom), this comparison is not necessarily a valid one according to the above theory, since certainly $k > 0$.

(6) AUTOCORRELATION AND AUTOREGRESSIVE ANALYSIS OF TIME-SERIES.

The classical method of analysing the structure of stationary time-series is that of harmonic analysis, but in a pioneering paper an alternative approach was suggested by Yule (1927), by way of the auto-regressive or autocorrelative structure of the series. He discussed in particular the analysis of the Wolfer sunspot series by such methods, using the simple autoregressive scheme

$$X_{t+2} + aX_{t+1} + bX_t = W_{t+2} \tag{28}$$

where W_{t+2} is independent of (or at least uncorrelated with) $W_{t+2-\tau}$, $(\tau > 0)$, and of constant variance σ^2. Similar schemes have since been used as approximate models for many time-series (see, for example, M.G. Kendall (1946b)). To examine the status of such representations in relation to the general specification problem, let us write (28) in the (backward) operational form

$$H_t X_{t+2} \equiv (1+aE^{-1}+bE^{-2})X_{t+2} = W_{t+2} \tag{29}$$

where $E \equiv 1+\Delta$, (this displacement operator is not to be confused with the expectation operator E). The solution of (29) is

$$X_t = H_t^{-1} W_t = \sum_{-\infty}^{t} g_{t-u} W_u, \tag{30}$$

where for (29)

$$g_u = (\mu_1^{1+u} - \mu_2^{1+u})/(\mu_1 - \mu_2),$$

μ_1 and μ_2 being the roots (assumed with modulus less than one) of $\mu^2+a\,\mu+b = 0$. This suggests that we consider the linear representation (30) for more general H_t or g_u. It then includes other well-known special processes such as 'moving averages'. In its continuous time analogue it also includes what in physics are called 'shot effects' (see, for example, Rice, 1944-5), but it will be convenient to deal first with the discrete time case. It turns out that the representation (30), with W_u at least an uncorrelated sequence, is quite general for any stationary non-deterministic time-series. This incidentally provides an immediate solution to Wiener's problem of predicting $X_{t+\tau}$, for

$$X_{t+\tau} = \sum_{-\infty}^{t} g_{t+\tau-u} W_u + \sum_{t+1}^{t+\tau} g_{t+\tau-u} W_u, \tag{31}$$

where the first component is in theory predictable and the second, with variance

$$\sum_{1}^{\tau} g_u^* g_u \; \sigma^2 \tag{32}$$

depends on future W_u and is unpredictable*. Wiener's solution may thus interpreted as a general method for writing $X_{t+\tau}$ in the form (30).

* For independent W_u (and linearly unpredictable in other cases).

The autocovariance of X_t is easily seen to be

$$w_r = \sum_1^\infty g_u^* \, g_{u+r} \, \sigma^2, \tag{33}$$

and in particular its variance

$$\sigma^2(X) = \sum_1^\infty g_u^* \, g_u \, \sigma^2. \tag{34}$$

The corresponding spectrum is found to consist of a density function

$$\frac{dF(\omega)}{d\omega} = f(\omega) = \frac{h(\omega)h^*(\omega)\sigma^2}{2\pi\sigma^2(X)} \tag{35}$$

where $h(\omega)$ is the transform of g_u,

$$h(\omega) = \sum_0^\infty e^{-iu\omega} g_u. \tag{36}$$

These results tend to support the view that while in theory autocorrelation and harmonic analysis of time-series are merely two aspects of the same analysis, in practice it may be convenient to use autocorrelation and autoregressive analysis in particular for series with no exact harmonic and deterministic components but with essentially continuous spectra; for example, it may then be found possible, where the structure or mechanism of the process is not known, to represent it by some finite representation of this type, like (28).

Two inference problems then arise, as in the probability chain case, firstly, to estimate any unknown coefficients or constants in such a specification and secondly, to test the goodness of fit of the resulting model. A general and useful estimation method for these linear representations is the 'least-squares' method. For example, to estimate a and b in (28), we minimize the sum of squares

$$\sum_1^{n-2} W_{u+2}^2 = \sum_1^n (X_{u+2} + aX_{u+1} + bX_u)^2, \tag{37}$$

where it is assumed that W_u has zero mean, and hence correspondingly that X_u is also measured from its true mean m or from the arithmetic mean $\overline{X}$ ⁄ if m is unknown. Two points to notice about this method are (a) that if the distribution of W_u is assumed normal, X_t would be a normal process and the expression (37) would appear in the exponential of the likelihood function (b) W_u is equivalent to X_u, given X_{u-1} and X_{u-2}, and the likelihood function would be completed with the probability term for X_1, X_2 (cf. §8). This last term merely causes an 'end effect' which may be neglected in a long series. Thus the least-squares estimates as above defined are also asymptotically the maximum likelihood estimates for a normal process, but of course, as noted in §3, are also estimates with useful and known properties for a much wider class of series. It has in fact been shown that the properties which least-squares estimates have in classical statistics are asymptotically true for autoregressive least-squares estimates, at least if the W_u may be assumed independent (and with finite moments). The amount of bias that arises for short series, either due to the intrinsic properties of estimates obtained from the above series or due to other effects such as superposed error on our observed variables X_t, also require investigation, but will not be pursued further here.

In the goodness of fit problem it would be theoretically possible for any normal process to deduce a general asymptotic test by the principles outlined in §1, but it again is usually more practicable to concentrate on the correlational properties of autoregressive series, which may or may not be normal. Thus there will be available a set of observed autocorrelations r_s (or correlogram) estimated in the natural

---

* For independent W_u (and linearly unpredictable in other cases).

⁄ This estimate of m may be justified by including m specifically in (37)

$$C_s = \frac{1}{T-s} \sum_{u=1}^{T-s} X_u X_{u+s}, \quad (s \geq 0), \tag{38}$$

(X_t being still measured from the mean). Such estimates could again be shown to be asymptotically equivalent to maximum likelihood estimates for normal processes, but moreover have useful properties in other cases. They are, for example, consistent estimates under the general ergodic hypothesis, and their asymptotic standard errors have been investigated (see, for example, Bartlett, 1946). There will also be a set of theoretical autocorrelations ρ_s calculates on the basis of the auto-regressive model. For example, multiplying (28) by X_{t-s} and averaging, we may obtain the difference equation for ρ_s

$$\rho_{s+2} + a\rho_{s+1} + b\rho_s = 0, \quad (s = -1,0,1,\ldots). \tag{39}$$

A systematic method of comparing these two correlograms has been given by Quenouille (1947), who has shown that the linear combination $H_t^2 r_t$ is asymptotically uncorrelated and hence that a series of χ^2 components can be built up from the observed correlogram. The actual expression for each component is

$$(H_t^2 r_t)^2 (n-t)\sigma^4/\sigma^4(X), \quad (t = p+1,\ldots) \tag{40}$$

where p is the 'length' of the operator H (e.g. 2 for equation (28)). As a numerical illustration I quote Quenouille's results for such a check on an artificial series of 480 terms constructed by M.G. Kendall (1946b) from the scheme

$$X_{t+2} - 1.1X_{t+1} + 0.5X_t = Y_{t+2} \tag{41}$$

(with Y_t independent but with a rectangular distribution). For this series

$$H_t^2 r_t = (1-1.1E^{-1}+0.5E^{-2})^2 r_t, \quad (t=3,\ldots)$$

Table I gives the relevant comparison. It will be noted that the total χ^2 (15 degrees of freedom and hence expectation 15) is satisfactory, as would be anticipated for a series which is know a priori to be the correct model*.

Table I.

t	ρ_t	r_t	χ^2_{t+2}	Total χ^2
1	0.7333	0.762	2.32	2.32
2	0.3067	0.377	0.07	2.39
3	-0.0293	0.079	1.56	3.95
4	-0.1856	-0.067	0.26	4.21
5	-0.1895	-0.078	0.00	4.21
6	-0.1156	-0.039	1.62	5.83
7	-0.0325	-0.007	1.99	7.81
8	0.0221	0.022	3.31	11.12
9	0.0406	0.018	0.43	11.55
10	0.0336	-0.036	0.59	12.14
11	0.0166	-0.103	0.22	12.36
12	0.0015	-0.145	0.70	13.06
13	-0.0067	-0.128	0.13	13.19
14	-0.0081	-0.052	0.10	13.29
15	-0.0056	0.029	0.02	13.32

* For further discussion and extensions of this technique to cover other types of linear scheme (including processes defined for continuous time) see, for example, Bartlett and Diananda (1950). One curious result when estimation is necessary is that, whereas for the general goodness of fit of §1, the estimation must be efficient in the maximum likelihood sense, in the present correlational goodness of fit test, the estimation need not even be efficient in the least-squares sense, provided that the errors are of the latter order of smallness $(n^{-\frac{1}{2}})$.

'Continuous' processes. As was apparent in the specification of time-series summarized in §2, there is a close parallel between the theory of series defined for discrete and continuous time. The linear process corresponding to (30) may be written

$$X(t) = \int_{-\infty}^{t} g(t-u)dY(u) \tag{42}$$

and includes the solutions of linear differential equations containing a random impulsive term. For example, the mechanical problem for which Yule intended (28) to be a model, namely, a swinging pendulum bombarded with peas by pea-shooting boys (or, more sedately, a torsional galvanometer bombarded by gas molecules), should strictly be represented by the 'differential' equation (with $\dot{X} \equiv dX/dt$)

$$d\dot{X}(t) + \alpha\dot{X}(t)dt + \beta X(t)dt = dY(t), \tag{43}$$

where each impulse $dy(t)$ is independent (or at least uncorrelated) with previous impulses, with solution, obtained by standard methods, of the type (47). The autocovariance function of (42) is well-known from its use as a model for the 'shot-effect' to be

$$w(\tau) = \sigma^2 \int_0^{\infty} g^{*}(u)g(u+\tau)du \tag{44}$$

where σ^2 is the increase in variance per unit time of $Y(t)$. Similarly the spectral density $f(\omega)$ is still given by formula (35), defined now for ω from $-\infty$ to ∞, and with

$$h(\omega) = \int_0^{\infty} e^{-iu\omega} g(u)du. \tag{45}$$

The formal analogue of the least-squares estimation of a and b in (28) is for the series (43) the minimization of

$$\int_0^T \left[dX(t) + \alpha X(t)dt + \beta \dot{X}(t)dt \right]^2$$

leading to the solution

$$\left.\begin{aligned} \int_0^T \dot{X}(t) \left[d\dot{X}(t) + \alpha\dot{X}(t)dt + \beta X(t)dt \right] &= 0, \\ \int_0^T X(t) \left[d\dot{X}(t) + \alpha\dot{X}(t)dt + \beta X(t)dt \right] &= 0. \end{aligned}\right\} \tag{46}$$

Since

$$\int_0^T X(t)\dot{X}(t)dt = \left[\tfrac{1}{2}X(t) \right]_0^T$$

$$\int_0^T X(t)d\dot{X}(t) = \left[X(t)\dot{X}(t) \right]_0^T - \int_0^T \dot{X}^2(t)dt$$

the solutions of (46) are asymptotically equivalent to the estimates

$$\left.\begin{aligned} \alpha_e &= -\int_0^T \dot{X}(t)d\dot{X}(t) \Big/ \int_0^T \dot{X}^2(t)dt \\ \beta_e &= \int_0^T \dot{X}(t)dt \Big/ \int_0^T X^2(t)dt \end{aligned}\right\} \tag{47}$$

---

* For further discussion and extensions of this technique to cover other types of linear scheme (including processes defined for continuous time) see, for example, Bartlett and Diananda (1950). One curious result when estimation is necessary is that, whereas for the general goodness of fit of §1, the estimation must be efficient in the maximum likelihood sense, in the present correlational goodness of fit test, the estimation need not even be efficient in the least-squares sense, provided that the errors are of the latter order of smallness ($n^{-\frac{1}{2}}$).

and their asymptotic errors from least-squares theory are

$$\sigma^2(\alpha_e) \sim 2\alpha/T, \quad \text{covariance } (\alpha_e, \beta_e) \sim 0.$$
$$\sigma^2(\beta_e) \sim 2\alpha\beta/T, \tag{48}$$

For series where a continuous track is available from electrical or optical means it is possible that estimates of the above type could be measured directly. When arithmetical methods are used, it is more convenient to use the observed autocorrelations, but limiting formulae of the type (47) and (48) are still useful in suggesting consistent methods of estimation having maximum (least-squares) efficiency (and their validity is checked by such further investigation; see Bartlett, 1946). It may be noted further that the least-squares equation for $\underline{a}$ and $\underline{b}$ in (28) is asymptotically equivalent to the first two equations (s=1 and 0) of the relation (39). The solution (46) is similarly equivalent to the equation

$$\rho''(\tau) + \alpha\rho'(\tau) + \beta\rho(\tau) = 0, \quad (\tau > 0), \tag{49}$$

(where dashes denote differentation), evaluated at $\tau = 0$, and to its derivative, also evaluated at $\tau = 0$. If it is considered to use instead a relation like (39), the relevant equation is of the same form, with

$$a = -2e^{-\frac{1}{2}\alpha} \cos\sqrt{(\beta - \tfrac{1}{4}\alpha^2)}, \quad b = e^{-\alpha},$$

but holding only for s=0,1,... and not for s=-1.

With regard to the sampling errors of the observed autocorrelations, to remarks are sufficient. Firstly, if a discrete series of observations is used in place of the continuous record, the formula available for 'discrete' series will be appropriate; secondly, if the continuous record is used directly, the appropriate formulae are obtained immediately as limiting analogues of the discrete case. The overall goodness of fit test of theoretical correlations obtained from the autoregressive scheme can also be adapted to 'continuous' schemes of any type (42) (as mentioned in an earlier footnote), though this is less direct.

Autoregressive and other linear representations of time-series have been extended to more than one series, especially in econometrics. The problems raised are mainly similar in principle to those for a single series, but tend to be more complex and will not be considered in detail. Some care is necessary, especially in econometric applications, over the 'identification' problem, that is, whether the structural model is uniquely determined from the probability model.

(7) HARMONIC ANALYSIS OF TIME-SERIES*.

The classical periodogram analysis of a discrete series consisted of computing

$$A = \sqrt{\frac{2}{n}} \sum_{r=1}^{n} X_r \cos\frac{2\pi pr}{n}, \quad B = \sqrt{\frac{2}{n}} \sum_{r=1}^{n} X_r \sin\frac{2\pi pr}{n},$$

and hence the 'intensity' $I_p = A_p^2 + B_p^2$. The factor 2/n is arbitrary, but has been inserted to make the mean value of I_p equal to $2\sigma^2(X)$ for a completely random series. It is easily found that

$$I_p = 2 \sum_{s=-n+1}^{n-1} \left(1 - \frac{|s|}{n}\right) C_s \cos\omega s \tag{50}$$

where $\omega = 2\pi p/n$, $C_{-s} \equiv C_s$ /, and it is still assumed that E(X)=0. The average of equation (50) gives

$$E(I_p) = 2\sigma^2(X) \sum_{s=-n+1}^{n-1} \left(1 - \frac{|s|}{n}\right) \rho_s \cos\omega s \tag{51}$$

---

* Cf. Bartlett (1950).
/ Defined by equation (38)

and as n increases this gives in the limit the inverse relation between f_s and $f(\omega)$, namely,

$$2\pi\sigma^2(X)f(\omega) = \sum_{-\infty}^{\infty} \rho_s \cos \omega s. \qquad (52)$$

This emphasizes the point already made on the theoretical equivalence of autocorrelation and harmonic analysis, but it is important to notice that in practice it is the quantity (50) which will be available, not (51).

The classical assumption was that the spectrum of X_t was discrete, and the above method is then successful in isolating such harmonic components. It may readily be shown that if there is a discrete spectral component at Ω, then I_p is $O(n)$ at $\omega = \Omega$, but rapidly drops to $O(1/n)$ at frequency values $O(1/n)$ from Ω. Allowance was later made for observational errors (uniform 'noise'); it was recognized that the intensity I_p fluctuates about its mean value $2\sigma^2(X)$ in the absence of true periodicity for a completely random series, and such fluctuations, which follow the probability law

$$p(I_p \geqslant z) = e^{-z/E(I_p)}, \qquad (53)$$

at least for X_r normal, may be allowed for in assessing the significance of the periodogram peaks.

But this classical approach did not allow for the possibility of time-series with continuous spectra other than the uniform 'noise' case. A non-uniform continuous spectrum may have peaks comparable for a finite n with the value of $E(I_p)$ even at the frequency Ω of a discrete harmonic component, especially as the fluctuations in I_p possess the remarkable property, implicit in (53) in the case of the uniform spectrum, of being of the same order as the mean value $E(I_p)$, irrespective of the length of series available for analysis. Formula (53) holds in fact asymptotically for a wide class of series with continuous spectra; moreover, the correlation between neighbouring frequencies tends to zero.

Such series occur frequently in practice, and the neglect of the above properties has led in the past to many fallacious claims for significant "periods". The periodogram exhibits a wildly fluctuating appearance, and is useless without modification. Averaging over neighbouring values of ω was suggested by the late P.J. Daniell. This device is especially useful in automatic electrical or optical analysis of a continuous record (for which it may be shown that completely analogous properties hold), and indeed it has already implicitly been used, - for example, in electrical measurements of turbulence spectra.

For arithmetical analysis the following smoothing procedure has been suggested. The same formula (50) is used, but with a different interpretation. The total length of series is now nm, and C_s refers to the covariance obtained from the whole series. The formula then has the asymptotic effect of averaging intensities obtained from m series of length n. The resolving power depends on n the smoothing depends on m, fluctuations being of the order $1/\sqrt{m}$ their unsmoothed value. As an illustration consider again the artificial series (40), for which M.G. Kendall has demonstrated the futility of calculating the (unsmoothed) periodogram. The true spectrum, obtained from formula (35), is

$$2\pi f(\omega) = \frac{(1-b)(1-a^2+b^2+2b)}{(1+b)\left\{1+a^2+b^2-2b+2a(1+b\cos\omega+4b^2\cos\omega\right\}}, \quad (-\pi \leqslant \omega \leqslant \pi), \qquad (54)$$

and in Table II this is compared with the observed spectrum (smoothed with $n = 30$, $m = 16$).

Table II ($2\pi f(\omega)$ tabulated against $q = 30\omega/\pi$)

q	True Values	Observed	q	True Values	Observed
1	2.219	2.938	12	0.549	0.687
2	2.380	3.254	14	0.305	0.343
3	2.656	3.517	16	0.190	0.199
4	3.030	2.905	18	0.130	0.123
5	3.395	2.294	20	0.096	0.143
6	3.493	2.477	22	0.076	0.111
7	3.091	2.792	24	0.064	0.070
8	2.353	2.371	26	0.056	0.058
9	1.641	1.247	28	0.053	0.057
10	1.118	0.887	30	0.051	0.058

Of course, it will sometimes not be necessary to deal directly with the spectrum in this way, as an autoregressive or autocorrelation analysis may prove practically more useful; but a direct empirical study of the spectrum will often be required. It is of considerable interest, for example, in the case of Wolfer's sunspot numbers, the series originally examined by Yule; and in the paper already referred to (from which the figures in Table II are taken) the smoothed spectrum of the sunspot numbers is also given.

In practice it will usually be clear a priori whether the specification should cover only discrete or continuous spectral components, or both. The last case is of course even more complicated, but in the case of a long series it should be possible to separate the continuous from the discrete components by examining the variation from one portion to another of the total series (consideration of a single fixed length is insufficient, because for any finite length a discrete component will give a finite amplitude not distinguishable from a continuous spectrum).

(8) USE OF THE LIKELIHOOD FUNCTION

For reasons already stressed, the estimation methods used in §6, in the autoregressive and autocorrelation analysis of time-series, have been in line with the linear and quadratic analysis of the last two sections, and are not dependent on a complete specification of the likelihood function. In conclusion I shall consider briefly the estimation problem for two examples* of continuous series from the more precise standpoint indicated in §3. First as an example of the type of representation (i) of §3 for continuous processes, consider the stationary normal Markoff process

$$X(t_{i+1}) - m = \rho_i \left[X(t_i) - m\right] + Y(t_i), \quad (t_{i+1} > t_i), \qquad (55)$$

where $\rho_i = \exp\left[-\mu(t_{i+1} - t_i)\right]$ and $Y(t_i)$ is normal with zero mean. (This process is a special case of the autoregressive schemes considered in §6, and in differential form is defined by

$$dX(t) + \mu X(t)dt = dY(t) \qquad (56)$$

where $Y(t)$ is a normal additive homogeneous process). To obtain the precise likelihood function, some care is necessary with the specification. For most practical applications it seems most relevant to assumed (as in §6) that the underlying variance of the disturbances $Y(t)$ which are

* Grenander (1950) considered only the estimation of the mean m for these examples, and my discussion here is somewhat wider.

maintaining stationarity of the series is constant, say σ^2 per unit time. Then for small constant time-intervals t we have in (55)

$$\sigma^2(Y) \sim \sigma^2 \Delta t, \quad 2\mu\sigma^2(X) = \sigma^2.$$

Hence for the sequence of ordered observations $S_n \equiv x_1, x_2, \ldots x_n$

$$f(S_n/m, \mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma(X)\prod_1^{n-1}\sigma(Y_i)} \exp\left\{-\frac{(x_1 - m)^2}{2\sigma^2(X)} - \sum_1^{n-1}\frac{[x_{i+1} - \rho_i x_i - m(1-\rho_i)]^2}{2\sigma^2(Y_i)}\right\}$$

and as $\Delta t \longrightarrow 0$

$$L(m, \mu, \sigma^2) - L(0, 1, \sigma^2) \longrightarrow$$

$$-\frac{\mu[x(o) - m]^2}{\sigma^2} + x^2(o) + \frac{1}{2\sigma^2}\left[2\mu\int_0^T (x_t - m)dx_t - 2\int_0^T (x_t - m)^2 dt\right]$$

$$-\frac{1}{2\sigma^2}\left[2\int_0^T x_t dt - \int_0^T x_t^2\, dt\right] + \tfrac{1}{2}\log\mu. \qquad (57)$$

Thus

$$\frac{\partial L}{\partial m} = \frac{(2+\mu T)}{\sigma^2}\left[\frac{x(0) + x(T) + \mu\int_0^T x_t dt}{2 + \mu T} - m\right] \qquad (58)$$

and

$$\frac{\partial L}{\partial m} = \frac{1}{2\mu} - \frac{[x(0) - m]^2}{\sigma^2} + \frac{\int_0^T (x_t - m)dx_t}{\sigma^2} - \frac{\mu}{\sigma^2}\int_0^T (x_t - m)^2 dt. \qquad (59)$$

From equation (58), the unbiased estimate of m, if μ is known, is

$$\frac{x(0) + x(T) + \mu\int_0^T x_t dt}{2 + \mu T} \qquad (60)$$

and its exact optimum variance is $\sigma^2/[\mu(2 + \mu T)]$, as may easily be verified. For large T the estimate becomes asymptotically the ergodic time-mean $\int_0^T x_t dt/T$ with asymptotic variance $\sigma^2/(\mu^2 T)$. If μ is not known, it has also to be estimated from (59), and since the pair of equations are no longer in the required form for minimum variance, no advantage over asymptotic estimates is necessarily gained. It will be seen further that the precise estimate of μ in (59) requires a knowledge of σ^2. However, for large T the first two terms may be neglected and the asymptotic estimate

$$\int_0^T (x_t - m)dx_t \Big/ \int_0^T (x_t - m)^2 dt \qquad (61)$$

obtained (in which, if m is unknown, is substituted its asymptotic estimate). The estimate (61) is identical with the asymptotic least-squares estimate, as it should be, and its asymptotic variance is $1/I(\mu) \sim 2\mu/T$.

The estimation of σ^2 has purposely not been considered in the above equations, for this would lead to difficulties which are strictly due to the above precise formulation becoming unrealistic in the limit as $\Delta t \longrightarrow 0$. The term

$$\operatorname*{Lt}_{\Delta t \longrightarrow 0} \Sigma (\Delta x)^2/T$$

would arise, which as $(\Delta x)^2$ is of order Δt, is of finite order but could hardly be evaluated in practice. A similar situation arises in the model for ordinary Brownian motion, where a variance estimate could theoretically be based on such a limit and be determined exactly for any finite period (being based on an infinite number of degrees of freedom). In the Markoff model the simple though inefficient estimate $\int_0^T (x_t - m)^2 dt/(2\mu T)$ will usually be adequate.

As the expression in (60) is (for μ unknown) the unbiased estimate with minimum variance when $T(t_i)$ is normal, it will also be the linear estimate with similar properties for any stationary process with the same autocorrelation function $\exp(-\mu \Delta t)$. But for a non-normal process for which the exact likelihood function is known, it may be possible to find a better estimate, as the following example will demonstrate. A simple Poisson process with rate μ (i.e. events occur at random in time, so that the total number occurring in any interval follows a Poisson distribution) has associated stationary normal variables Z_r at each occurrence t_r, and the series $X(t)$ is defined as $Z_0 (0 \leqslant t < t_1)$, $Z_1 (t_1 \leqslant t < t_2)$, etc.* Since the contribution to the autocorrelation is 1 with probability $e^{-\mu \Delta t}$ and 0 with probability $1 - e^{-\mu \Delta t}$, this process has the same autocorrelation as the preceding Markoff process (55). Its distribution is moreover normal at each point t, but it is not a normal process in the full continuous time sense, as may be seen from its likelihood function.

Let the number of occurrences in $0 \text{——} T$ be N, a Poisson variable with mean μT. When N is given as n, say, the distribution of the actual times of occurrence $t_1, t_2, \ldots t_n$ is uniform in $0 \text{——} T$, and does not provide any information on m or μ. Using the second type of representation (ii) of §3, we have

$$f(m, \mu) = \frac{(\mu T)^n e^{-\mu T}}{n.'} \cdot \frac{\exp - \frac{1}{2\sigma_z^2} \sum_0^n (z_r - m)^2}{(2\pi)^{\frac{1}{2}(n+1)} \sigma_z^{n+1}},$$

where σ_z^2 is the variance of Z. Thus

$$L(m, \mu) - L(0, 1)$$
$$= n \log \mu - (\mu - 1)T + \frac{m}{\sigma_z^2} \sum_0^n z_r - \tfrac{1}{2} \frac{(n+1)m^2}{\sigma_z^2}, \qquad (62)$$

and

$$\frac{\partial L}{\partial m} = \frac{1}{\sigma_z^2} \left\{ \sum_0^n x_r - (n+1)\, m \right\}, \qquad (63)$$

$$\frac{\partial L}{\partial \mu} = \frac{n}{\mu} - T. \qquad (64)$$

---

* Grenander notes that this process has been used in the theory of servo-mechanisms.

Equation (63) is not of the required form to provide an unbiased estimate of m with minimum variance (n being variable). However, the maximum likelihood estimate of m is

$$\hat{m} = \frac{1}{n+1} \sum_{0}^{n} x_r \qquad (65)$$

with asymptotic variance

$$\frac{1}{I(m)} = \frac{\sigma_z^2}{E(n+1)} = \frac{\sigma_z^2}{1+\mu T} \sim \frac{\sigma_z^2}{\mu T} \quad , \qquad (66)$$

(the exact variance of $\hat{m}$ is also easily evaluated if required). For this process $\sigma^2(X) = \sigma_z^2$, and the linear estimate $\int_0^T x_t dt/T$ of the mean, with asymptotic variance $2\sigma^2(X)/(\mu T)$, has a limiting efficiency, measured by the ratio of these error variances, of $\frac{1}{2}$. It should be noticed that (65) is not a linear estimate in the sense previously defined, for, regarded as an integral for x_t, the weights depend on the realisation.

For μ, equation (64) gives $\hat{\mu} = n/T$, with variance μ/T. The "least-squares" estimate for μ is not applicable here, as it is $\sigma^2(X) = \sigma_z^2$ which is given constant. Moreover, no difficulty would exist in the estimation of σ_z^2; the natural estimate $\sum_{0}^{n} (z_r - \hat{m})^2/(n+1)$, (or divisor n if the estimate is to be unbiased), would be deduced, the only special feature of (62) compared with a sample of n independent normal variables being the random nature of n.

REFERENCES

Bartlett, M.S.	J. Roy. Statist.Soc.Suppl. 8, 27, (1946)
Ibid	Biometrika, 37, 1, (1950)
Ibid	Proc.Camb.Phil.Soc. (in Press)
Bartlett, M.S., and Diananda, P.H.	J.R.Statist.Soc. (in Press)
Chandrasekhar, S.	Reviews of Modern Physics, 15, 1, (1943)
Cramer, H.	"Mathematical Methods of Statistics", (Princeton). (1946)
Fisher, R.A.	"Statistical Methods for Research Workers", (Edinburgh). (1925)
Good, I.J.	"Probability and the Weighing of Evidence", (London). (1950)
Grenander, U.	Arkiv for Matematik, Band 1, nr.17. (1950)
Kendall, M.G.	"Advanced Theory of Statistics", Vol. 2, (London). (1946a)
Kendall, M.G.	"Contributions to the Study of Oscillatory Time Series", (Cambridge). (1946b)
Levy, P.	"Processus Stochastiques et Mouvement Brownien" (Paris). (1948)

Moyal, J.E.	J.R.Statist.Soc. Series B, 11, 150, (1949)
Quenouille, M.H.	J.R.Statist.Soc. 110, 123, (1947)
Rice, S.O.	Bell System Techn.J. 23, 282 and 24, 46, (1944 - 5)
Shannon, C.	Bell System Techn.J. 27, 379 and 623, (1948)
Wiener, N.	"Cybernetics" (New York). (1948)
Ibid	"Extrapolation, Interpolation and Smoothing of Stationary Time Series" (New York). (1949)
Yule, G.U.	Phil. Trans. A 226, 267, (1927)

GENERAL TREATMENT OF THE PROBLEM OF CODING

by
C.E. Shannon

ABSTRACT

A typical communication system consists of the following five elements:

(1) An information source. This can be considered to be represented mathematically by a suitable stochastic process which chooses one message from a set of possible messages. The rate R of producing information is measured by the entropy per symbol of the process.

(2) An encoding or transmitting element. Mathematically this amounts to a transformation applied to the message to produce the signal, i.e., the encoded message.

(3) A channel on which the signal is transmitted from transmitter to receiver. During transmission the signal may be perturbed by noise.

(4) A receiving and decoding (or demodulating) device which recovers the original message from the received signal.

(5) The destination of the information, e.g., the human ear (for telephony) or the eye (for television). The characteristics of the destination may determine the significant elements of the information to be transmitted. For example, with sound transmission, precise recovery of the phases of components is not required because of the insensitivity of the ear to this type of distortion.

The central problems to be considered are how one can measure the capacity of a channel for transmitting information; how this capacity depends on various parameters such as bandwidth, available transmitter power and type of noise; and what is the best encoding system for a given information source to utilize a channel most efficiently.

Since the output of any information source can be encoded into binary digits using, statistically, R binary digits per symbol, the problem of defining a channel capacity can be reduced to the problem of determining the maximum number of binary digits that can be transmitted per second over the channel.

When there is no noise in the channel, it is generally possible to set up a difference equation whose asymptotic solution gives essentially the number of different signals of duration T when T is large. From this, it is possible to calculate the number of binary digits that can be transmitted in time T and, consequently, the channel capacity.

In a noisy system, the problem is mathematically considerably more difficult. Nevertheless, a definite channel capacity C exists in the following sense. It is possible by proper encoding of binary digits into allowable signal functions to transmit as closely as desired to the rate C binary digits per second with arbitrarily small frequency of errors. There is no method of encoding which transmits a larger number. In general, the ideal rate C can only be approached by using more and more complex encoding systems and longer and longer delays at both transmitter and receiver.

The channel capacity C is given by an expression involving the difference of two entropies. This expression must be maximized over all possible stochastic processes which might be used to generate signal functions.

The actual numerical evaluation of C is difficult and has been carried out in only a few cases. Even when C is known, the construction of coding systems which approach the ideal rate of transmission is often unfeasible.

A simple example of a noisy channel in which the capacity and an explicit ideal code can be found is the following. Assume the elementary signals are binary digits and that the noise produces at most one error in a group of seven of these. The channel capacity can be calculated as 4/7 bits per elementary signal. A code which transmits at this rate on the average is as follows. Let a block of seven symbols be x_1, x_2, x_3, x_4, x_5, x_6, x_7 (each x either 0 or 1). x_3, x_5, x_6 and x_7 are used as message symbols, and x_1, x_2 and x_4 are used redundantly for checking purposes. These are chosen by the following rules:

(1) x_4 is chosen so that $\alpha = (x_4 + x_5 + x_6 + x_7) = 0 \text{ Mod } 2$

(2) x_2 is chosen so that $\beta = (x_2 + x_3 + x_5 + x_7) = 0 \text{ Mod } 2$

(3) x_1 is chosen so that $\gamma = (x_1 + x_3 + x_5 + x_7) = 0 \text{ Mod } 2$.

The binary number $\alpha\beta\gamma$, calculated by these same expressions from the received signal, gives the location of the error. (If zero, there was no error.) This forms a completely self-correcting code for the assumed type of noise.

If the signal functions are capable of continuous variation we have a continuous channel. If there were no noise whatever, a continuous channel would have an infinite capacity. Physically, there is always some noise. With white Gaussian noise the capacity is given by

$$C = W \log \left(1 + \frac{P}{N}\right) \tag{1}$$

in which

W = bandwidth in cycles per second

P = available average transmitter power

N = average noise power within the band W.

The equation (1) is an exchange relation among the quantities W, P, N and C. Thus, the transmitter power can be reduced by increasing the bandwidth, retaining the same channel capacity. Conversely a smaller bandwidth can be used at the expense of a greater signal-to-noise ratio.

If, as is usually the case, the noise power increases proportionally with bandwidth, $N = N_o W$, we have

$$C = W \log \left(1 + \frac{P}{N_o W}\right). \tag{2}$$

As W increases, C approaches the asymptotic value

$$C_\infty = \frac{P}{N_o} \log e \,. \tag{3}$$

If the perturbing noise is Gaussian but does not have a flat spectrum, the most efficient use of the band occurs when the sum of the transmitter power and the noise power at each frequency is constant,

$$P(\omega) + N(\omega) = K. \tag{4}$$

When the noise is Gaussian, it turns out that most efficient coding requires that the transmitted signal have the same statistical structure as Gaussian noise.

If the perturbing noise is not Gaussian, the mathematical problems of calculating channel capacity and ideal codes are formidable. The most that is known for the general case are upper and lower bounds for the channel capacity given by the following inequalities

$$W \log \left(\frac{P + N_1}{N_1}\right) \leq C \leq W \log \left(\frac{P + N}{N_1}\right) ,$$

where P, N and C are as before, and N_1 is the average power in a thermal noise having the same entropy as the actual noise. N_1 is a measure of the amount of randomness in the noise. It is intuitively reasonable that this should be a controlling term in the channel capacity since the more predictable the noise the more it can be compensated.

Among communication systems in actual use PCM (Pulse Code Modulation) and PPM (Pulse Position Modulation) come reasonably close to the ideal limits of channel capacity with white Gaussian noise. For high signal-to-noise ratio PCM is most appropriate. When the number of quantized amplitude levels is suitably adjusted, this method of modulation requires some eight to ten db greater power than the theoretical minimum. With low signal-to-noise ratios, PPM requires about the same extra signal power except for extremely low P/N values, in which case it is still closer to the ideal. Other more involved codes have been investigated, although not yet put into practice, which are about two db closer than PCM to the ideal. Rice has shown that certain types of codes approach the ideal roughly according to $1/\sqrt{\tau}$ where τ is the delay involved in the encoding process.

The general principles of communication theory and coding have an application in the study of secrecy systems. A secrecy system can be considered to be a communication system in which the noise is the arbitrariness introduced by the encoding process. It can be shown under certain assumptions that the redundancy of the original language is the fundamental factor governing the amount of material that must be intercepted in order to solve a cipher. These results check reasonably well against experimentally known results for certain simple secrecy systems.

REFERENCES:

1. Shannon, C.E. and Weaver, W. "The Mathematical Theory of Communication", University of Illinois Press, Urbana, 1949.

2. Shannon, C.E. "Communication Theory of Secrecy Systems", Bell System Technical Journal, vol.28, pp. 656-715, October 1949.

THE LATTICE THEORY OF INFORMATION

by

C.E. Shannon

ABSTRACT

The word "information" has been given many different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field. The present note outlines a new approach to information theory which is aimed specifically at the analysis of certain communication problems in which there exist a number of information sources simultaneously in operation. A typical example is that of a simple communication channel with a feedback path from the receiving point to the transmitting point. The problem is to make use of the feedback information for improving forward transmission, and to determine the forward channel capacity when the best possible use is made of this feedback information. Another more general problem is that of a communication system consisting of a large number of transmi ting and receiving points with some type of interconnecting network between the various points. The problem here is to formulate the best systems design whereby, in some sense, the best overall use of the available facilities is made. While the analysis sketched here has not yet proceeded to the point of a complete solution of these problems, partial answers have been found and it is believed that a complete solution may be possible.

(1) THE NATURE OF INFORMATION

In communication theory we consider information to be produced by a suitable stochastic process. We consider here only the discrete case; the successive symbols of the message are chosen from a finite "alphabet", and it is assumed for mathematical simplicity that the stochastic process producing the message has only a finite number of possible internal states. The message itself is then a discrete time series which is one sample from the ensemble of possible messages that might have been produced by the information source. The entropy $H(\alpha)$ of such a source is a measure of the amount of information produced by the source per letter of message. However, $H(\alpha)$ can hardly be said to represent the actual information. Thus, two entirely different sources might produce information at the same rate (same H) but certainly they are not producing the same information.

To define a concept of actual information, consider the following situation. Suppose a source is producing, say, English text. This may be translated or encoded into many other forms (e.g. Morse code) in such a way that it is possible to decode and recover the original. For most purposes of communication, any of these forms is equally good and may be considered to contain the same information. Giver any particular encoded form, any of the others may be obtained, (although of course it may require an involved computation to do so). Thus we are led to define the actual information of a stochastic process as that which is common to all stochastic processes which may be obtained from the original by reversible encoding operations. It is desirable from a practical standpoint and mathematically convenient to limit the kind of all allowed encoding operations in certain ways. In particular, it is desirable to require that the encoding be done by a transducer with a finite number of possible internal states. This finite memory condition prevents paradoxical situations in which information goes into a transducer more rapidly on the average than it comes out at the output.

Each encoded version of the original process may be called a translation of the original language. These translations may be viewed as different ways of describing the same information in about the same way that a vector may be described by its components in various coordinate systems. The information itself may be regarded as the equivalence class of all translations or ways of describing the same information.

(2) THE METRIC, TOPOLOGY AND CONVERGENT SEQUENCES

With this definition of information, it is possible to set up a metric satisfying the usual requirements. The metric $\rho(x,y)$ measures the distance between two information elements x and y, and is given in terms of conditional entropies. We define

$$\rho(x,y) = H_x(y) + H_y(x) = 2\,H(x,y) - H(x) - H(y).$$

The symmetry property $\rho(x,y) = \rho(y,x)$ is obvious from the definition. If $\rho(x,y) = 0$, both $H_x(y)$ and $H_y(x)$ must be zero (since both are necessarily non-negative), and this requires that the x sequence be calculable with probability 1 from the y sequence and vice versa. The triangle law for a metric

$$\rho(x,y) + \rho(y,z) \geq \rho(x,z)$$

is readily shown by expanding these terms into the various entropies and making use of known inequalities for entropies. It may be noted that $\rho(x,y)$ is independent of the particular translations of x and y used in its calculation. This is due to the fact that $H_x(y)$ and $H_y(x)$ are invariant under finite state encoding operations applied to x and y.

The existence of a natural metric enables us to define a topology for a set of information elements and in particular the notion of sequences of such elements which approach a limit. A set of information elements $x_1, x_2, \ldots, x_n, \ldots$ will be said to be Cauchy convergent if

$$\lim_{\substack{m \to \infty \\ n \to \infty}} \rho(x_m, x_n) = 0$$

The introduction of these sequences as new elements (analogous to irrational numbers) completes the space in a satisfactory way and enables one to simplify the statement of various results.

(3) THE INFORMATION LATTICE

A relation of inclusion, $x \geq y$, between two information elements x and y can be defined by

$$x \geq y \equiv H_x(y) = 0$$

This essentially requires that y can be obtained by a suitable finite state operation (or limit of such operations) on x. If $x \geq y$ we call y an abstraction of x. If $x \geq y$, $y \geq z$, then $x \geq z$. If $x \geq y$, then $H(x) \geq H(y)$. Also $x > y$ means $x \geq y$, $x \neq y$. The information element, one of whose translations is the process which always produces the same symbol, is the 0 element, and $x \geq 0$ for any x.

The sum of two information elements, $z = x + y$, is the process one of whose translations consists of the ordered pairs (x_n, y_n) where x_n is the nth symbol produced by the x sequence and similarly for y_n. We have $z \geq x$, $z \geq y$ and there is no $w < z$ with these properties; z is the least upper bound of x and y. The element z represents the total information of both x and y.

The product $z = xy$ is defined as the largest z such that $z \leq x$, $z \leq y$; that is, there is no $w > z$ which is any abstraction of both x and y. The product is unique. Here z is the common information of x and y.

With these definitions a set of information elements with all their sums and products form a metric lattice. The lattices obtained in this way are not, in general, distributive, nor even modular. However they can be made to be relatively complemented by the addition of suitable elements. For $x \leq y$ it is possible to construct an element z with

$$z + x = y$$
$$zx = 0 .$$

The element z is not, in general, unique.

The lattices obtained from a finite set of information sources are of a rather general type; they are at least as general as the class of finite partition lattices. With any finite partition lattice it is possible to construct an information lattice which is abstractly isomorphic by a simple procedure.

Some examples of simple information lattices are shown in Figs. 1 and 2.

In Fig. 1, there are three independent sources. The product of any two of these elements is zero, and the conventional lattice diagram is that shown at the right. In Fig. 2, there are two independent sources of binary digits, x and y. The sequence z is the sum mod 2 of corresponding symbols from x and y. In this case again the product of any two of x, y and z is zero, but the sum of any two represents the total information in the system. In this case the lattice is non-distributive, since $zy + zx = 0 + 0 = 0$, while $z(x+y) = z \neq 0$.

(4) THE DELAY FREE GROUP G1

The definition of equality for information based on the group G of all reversible encoding operations allows $x = y$ when y is, for example, a delayed version of x; $y_n = x_{n+a}$. In some situations, when one must act on information at a certain time, a delay is not permissible. In such a case we may consider the more restricted group G1 of instantaneously reversible translations. One may define inclusion, sum, product, etc., in an analogous way, and this also leads to a lattice but of much greater complexity and with many different invariants.

FIG. 1.

FIG. 2.

THEORY OF RADAR INFORMATION

by

P.M. Woodward

Radar is a system of measurement rather than communication, yet it is quite possible to apply information theory to it, in order to see whether the very small received signals inherently contain as much information as those of an ideal communication system working at the same signal-to-noise ratio. It turns out that they do, very nearly, but this is not really what makes radar a suitable topic for this symposium. The main interest is in the type of coding it represents. Shannon has pointed out [1] that when the natural number of dimensions of a message is artificially increased by mapping non-topologically into a signal space of higher dimensions, a marked threshold effect is produced. Radar exhibits such a threshold particularly well and it is to this that I wish to direct attention.

We shall consider only the most obvious radar problem, that of measuring the range of a stationary target. This is one-dimensional information, and it is important to realize that the way in which it is coded is almost entirely beyond control: it is determined by the very nature of radar. A known periodic waveform is transmitted, echoed, and received again. The time τ which elapses between transmission and reception represents the range of the target. Unlike most systems of electrical signalling, the choice of transmitted waveform does not represent the required information but forms part of the observer's a priori knowledge. All the required information is embodied in the time-delay of the received waveform. If we fix the received signal energy, this leaves the received signal with only one degree of freedom, but the noise which goes with it will have many degrees of freedom. In geometrical language, the received signal is treated as a point in a multi-dimensional waveform space, and the ensemble of received signals lie along a one-dimensional twisted curve embedded in this hyper-space, and incidentally lying on the surface of a hyper-sphere. One therefore expects a threshold effect as soon as the noise perturbation is sufficiently large to short-cut the convolutions of this message locus, and introduce wild ambiguities of range measurement.

However, there is really no difference in principle between interpreting a radar signal and any other kind of signal. The observer always has some a priori knowledge of what he is trying to receive, and it merely happens that in radar he knows, apart from noise, the exact shape of the waveform. Obviously his first step is, effectively, to place the transmitted and received waveforms side by side, and try to estimate the time-shift τ. If the possible values of τ are known a priori to be continuously distributed, it should be clear that in the presence of noise he cannot hope to determine τ exactly, because this would represent an infinite quantity of information and would require an infinite amount of signal energy. His best estimate is therefore bound to be subject to error. One might suppose, then, that the first problem is to determine theoretically the spread of his best guesses over an ensemble of received waveforms all resulting from the same true value of τ. The average quantity of information obtained in any determination would then be given by

$$I = H(y) - H_x(y).$$

Shannon [2] uses x to symbolise transmission and y to symbolise reception, but it must be understood that the "transmitted message" in radar has nothing to do with the radar transmitter; it refers to the true value of τ, while y refers to the observer's estimate of τ. The entropy $H(y)$ is a measure of the spread of variability of all previous guesses over a complete ensemble of true values, and therefore represents the a priori uncertainty about the next one, while the conditional entropy $H_x(y)$ represents the spread of the guesses in an

ensemble in which the true value is fixed and only the noise-sample is different. This might seem to be the obvious approach, but it is not altogether satisfactory because it appears to contain subjectives elements. It appears to depend on the particular way in which the observer makes his guess. If he were no good at guessing, $H_x(y)$ would be large and the quantity of information small. The maximum quantity of information latent in the received waveform could only be evaluated by this method by giving rules for making the best possible guess.

This whole difficulty is avoided by starting from the alternative formula.

$$I = H(x) - H_y(x)$$

Here $H(x)$ is the entropy of the a priori distribution for the true value of τ, not the guessed value. The equivocation, $H_y(x)$, represents the observer's uncertainty about the true value of τ on any one occasion, i.e. when the received waveform is fixed. We have, then, to consider a fixed received waveform, arising from a true range τ_0 say, and find out from it, not simply the most probable value of τ it might represent, but a complete probability distribution for all possible values of τ. This is not subjective at all: it represents the matter-of-fact frequency distribution of those values of τ which could have given rise to this particular τ_0-waveform. The whole problem thus centres round two distributions, the a priori distribution, called $p_0(\tau)$, and the a posteriori distribution denoted $p_1(\tau)$. The difference of the two corresponding entropies is the quantity of information gained.

Without entering into too much mathematical analysis, we may indicate in outline how $p_1(\tau)$ is found, because this is really the heart of the problem. Let us suppose that the received waveform is observed for a duration of time D, and denote it by

$$y(t) = u(t - \tau_0) + n(t),$$

where $u(t)$ is what would have been received with no time-delay τ_0 and no noise $n(t)$, and is presumed known to the observer. If he supposes the true value of the range to be τ, he calculates that the noise would have to be

$$y(t) - u(t - \tau).$$

The probability of such noise then determines inversely[3] the probability that his hypothesis τ was correct. Now the probability density for the random noise fluctuation in its multi-dimensional waveform space is proportional to

$$e^{-W/N_0}$$

where W is the total energy of the noise over the interval D, and N_0 is the mean noise power per unit bandwidth, so the probability density in favour of the hypothesis τ is given by

$$p_1(\ ) = \lambda \exp\left[-\frac{1}{N_0}\int_D \{y(t) - u(t - \ )\}^2 \, dt \right]$$

if the a priori probability distribution $p_0(\tau)$ is uniform. We shall in fact take $p_0(\tau)$ to be uniform over an arbitrary interval of range T, within which it is assumed that τ is known to lie. (This a priori interval T may be less than or equal to the repetition period of $u(t)$.) Consequently, $p_1(\tau)$, is only defined over the interval T and λ must be chosen to normalize it accordingly. We may note in passing that the most probable value of τ is that which gives the least mean square departure of y from $u(t - \tau)$.

Any factors in the expression for $p_1(\tau)$ which do not depend on τ may obviously be absorbed into the normalizing constant, and if D is an integral multiple of the repetition period, we are left only with

$$p_1(\tau) = \lambda \exp\left[\frac{2}{N_0}\int_D y(t)\, u(t - \tau)\, dt \right]$$

This expression is interesting because the integral in it has the familiar form of the output from a linear filter whose impulse response is $u(-t)$, the time-reverse of the transmitted waveform, and whose input is simply the received waveform $y(t)$. The limits of integration for such a filter, however, would go from t-D to t, whereas the limits above are quite fixed. To follow up this question would take too long, but it does have an interesting bearing on the topic of optimum filtering, and especially on the result of Van Vleck and Middleton [4] that just such a filter would give the maximum peak signal-to-noise ratio.

Having decided on a form for $p_o(\tau)$ and derived an expression for $p_1(\tau)$, there would seem to be nothing to prevent us, in principle, from calculating the corresponding entropies H_o and H_1 forthwith. Indeed, H_o can be seen immediately to equal log T, but the calculation of H_1 takes much longer and cannot be obtained without first investigating the properties of $p_1(\tau)$. This distribution itself is, in any case, more important both theoretically and practically than its entropy, for while the entropy enables us to determine the quantity of information for comparison with Shannon, it leaves the interpretation of the information - or lack of it - entirely out of account. We may begin examining $p_1(\tau)$ by writing the full expression for the received waveform in place of $y(t)$ in the integral. Then

$$p_1(\tau) = \lambda e^{g(\tau) + h(\tau)}$$

where

$$g(\tau) = \frac{2}{N_o}\int_D u(t - \tau_o)\, u(t - \tau)\, dt$$

$$h(\tau) = \frac{2}{N_o}\int_D n(t)\, u(t - \tau)\, dt$$

It will be seen that $g(\tau)$ is obtained from the signal actually received, and $h(\tau)$ from the noise. An observer, it must be remembered, could form $p_1(\tau)$ after any one observation, but he could not of course determine $g(\tau)$ and $h(\tau)$ separately as we are doing.

Consider first the "signal function" $g(\tau)$. The waveform $u(t)$ which generates it, is a high-frequency function of t and this makes $g(\tau)$ a high-frequency function of τ, but the envelope of $g(\tau)$ is slowly varying (by comparison with the carrier) because it is controlled by the bandwidth of $u(t)$. For the present, let us forget the carrier in $g(\tau)$, to which we shall return at the end, and concentrate on the envelope alone. It is almost obvious and not difficult to show mathematically, that this has a maximum at $\tau = \tau_o$, where its value is $2E/N_o$, E being the total received signal energy. In fact, the envelope of $g(\tau)$ will have a single peak at τ_o and be negligible, if not precisely zero, elsewhere. This last is not a mathematical deduction, it is a statement applying to practical waveforms, whether amplitude or frequency modulated, and is the very feature by which the suitability of any waveform for radar may be judged. It ensures that the message-locus is well spread out in waveform space.

The "noise function" $h(\tau)$ has certain features in common with the signal function. It has, for example, a slowly varying envelope controlled by the bandwidth of $u(t)$, but it is a stationary random function of τ, and has all the characteristics of noise which has passed through a high-frequency pass band filter, except that the RMS value of its envelope is $2\sqrt{E/N_o}$, which happens to increase as the received signal energy increases.

We now have sufficient facts to discuss all that is of qualitative importance about the a posteriori distribution. It is clear, to start with, that $p_1(\tau)$ is partly a random function of τ, owing to the presence of $h(\tau)$. It may seem a confusing idea that a probability distribution should itself be random, but it is simply a matter of being clear about ensembles. The distribution $p_1(\tau)$ represents the frequency with which various ranges τ could give rise to the exact

waveform which we have privately stated to be due on this occasion to a range τ_o. It will obviously depend, to some extent, on the particular way in which the noise happened to act on this occasion. With a fixed true range τ_o, therefore, $p_1(\tau)$ will be different from one occasion to another, and it is just this randomness which $h(\tau)$ represents.

It should be clear that if $E \lesssim N_o$, there will be no marked accumulation of probability near τ_o, because in terms of envelopes the RMS value $2\sqrt{E/N_o}$ of $h(\tau)$ will exceed the peak value $2E/N_o$ of $g(\tau)$. Throughout the remainder of the theory, we are in fact forced to assume that

$$E \gg N_o$$

for purely mathematical reasons, but since it is a necessary condition for satisfactory observation, the assumption is not seriously embarrassing. This energy criterion is not in any way peculiar to radar, and is not connected with the threshold effect due to coding. When $E \gg N_o$, the peak value of g will be so large that, after normalization, the probability distribution $p_1(\tau)$ will be almost unaffected by the presence of $h(\tau)$, at least for most values of τ. Indeed, the whole of the peak in $g(\tau)$ except a small region immediately surrounding its apex will be similarly reduced, and expansion of the exponent about τ_o will yield a Gaussian distribution for $p_1(\tau)$. The standard deviation works out to be

$$\sigma = \frac{1}{\beta}\sqrt{\frac{N_o}{2E}}$$

where β^2 is the second moment of the power spectrum of the transmitted waveform about its centroid. The quantity β is the one really important parameter associated with the transmitted waveform, and is equal, apart from a constant factor, to the "effective bandwidth" adopted by Gabor (5). The standard deviation σ gives us a tentative measure of the accuracy with which τ may be determined. As would be expected, σ decreases with an increase in signal energy and also with an increase of transmitter bandwidth such as might arise from the use of shorter pulses.

Two things appear at first to be wrong with the above result: the first is that the a posteriori distribution is apparently centred exactly on τ_o. Since it is within the observer's power to determine $p_1(\tau)$ on any occasion, apparently he could look for the centre of the peak and so determine τ_o exactly, which would contradict the very uncertainty the distribution is supposed (inversely) to describe. It is, of course, the noise function $h(\tau)$ which removes this paradox and it can be shown that its effect is to disturb the position of the peak in a random manner and by just the right amount. Its width is largely unaffected by the presence of h, however, so that σ remains a valid measure of range error. The second objection arises the moment comparison is made with general communication theory,[2] which sets a limit of E/N_o natural units of information on any message of energy E. Yet if β is increased and E/N_o kept fixed, it would appear that σ can be made as small as desired. This would make H_1 as small as desired and corresponds to increasing the quantity of received information without limit, which would contradict Shannon's fundamental theorem. Again it is the presence of $h(\tau)$ which prohibits this. As β is increased, the autocorrelation interval in $h(\tau)$ is proportionately reduced and more statistically independent opportunities exist for $h(\tau)$ to produce a spurious peak well in excess of its RMS value, and therefore big enough to show up on the normalized a posteriori distribution. Every time $h(\tau)$ succeeds in doing this, a spurious Gaussian distribution is introduced in $p_1(\tau)$ and there comes a stage when there are so many of these that $g(\tau)$ might almost as well not be present. These are conditions of completely ambiguous reception, in which the accuracy of measurement might be high if the observer only knew which peak to select. The ambiguity operates in such a way as to reduce the quantity of information by just the amount required to bring it within the fundamental limit of E/N_o natural units.

The a posteriori distribution thus describes two quite different kinds of uncertainty of reception. First there is the small connected region of uncertainty in the neighbourhood of the true range. This must inevitably be present as long as the signal energy is finite, and is no cause for complaint. But in addition, when β is too large, there is a wild uncertainty, even though the received signal energy is large, which prevents an observer from knowing even approximately whereabouts in the interval T the true value of τ is to be found. We may call this effect non-topological error, because it arises from the non-topological mapping of a one-dimensional ensemble of messages into a multi-dimensional waveform space. This effect shows up one of the weaknesses of judging a communication system solely in terms of quantities of information. When non-topological error is present, the system is useless from a practical point of view and yet the mathematical quantity of information may be quite large. It is a question of intelligibility rather than information.

Intelligibility is a concept associated with meaning, and it is not to be expected that a general theory of it should be quite as straightforward as that of information-content. However, it may happen in particular problems that a quantitative assessment is possible. In radar, it is especially simple. We define unintelligibility by the non-topological ambiguity of reception, A, given by the area under $p_1(\tau)$ which lies nowhere near the true value. The Figure illustrates the dependence of A on log $T\beta$ and E/N_0. As remarked above, ambiguity increases with β, but it also depends on E/N_0 and is responsible for a threshold of intelligibility as the total received energy increases, as it would with increasing time of obervation. The threshold extends over one or two units of E/N_0 and occurs (very roughly) where

$$\log T\beta = E/N_0$$

Contours of information, I, are also shown. It will be seen that they behave in a markedly different manner on either side of the threshold. In the ambiguous region they would be strictly asymptotic to the fundamental limit $I = E/N_0$ if the difference of H_0 and H_1 had been evaluated in a straightforward manner. But in making a detailed calculation of I, we have in fact tampered with the a posteriori distribution by smoothing out the fine structure produced by the "carrier" in $g(\tau)$ and $h(\tau)$. The effect of the carrier is to make the signal peak in $p_1(\tau)$ consist, not of a single Gaussian distribution of standard deviation σ, mentioned before, but of a closely packed sequence of very narrow Gaussian distributions under a Gaussian envelope with this standard deviation. We have thought it best to remove the fine structure information by short-scale smoothing of $p_1(\tau)$ because it is of no value in practice. It arises from the comparison which the observer could make between the carrier phases of transmitted and received waveforms, and its removal increases the a posteriori entropy by a term in log (E/N_0), which is just what prevents the absolute limit being attained in the ambiguous region. In the unambiguous region, the information I is limited solely by the topological error in range and increases comparatively slowly with energy. This is because additional energy, once the threshold has been crossed, is not employed in the systematic removal of ambiguity but in the improvement of range accuracy by continued repetition of information already partly known.

To sum up, it is seen that there are in radar two quite different conditions of reception. There is ambiguous reception in which the information rate is high and the intelligibility low, and there is unambiguous reception in which the information rate is low but intelligibility is high. It would therefore appear that merely to evaluate quantities of information, and compare them with the ideal limits is not adequate guide to the behaviour of a communication system. Some figure for intelligibility may also be necessary in particular problems. The analysis of radar shows that loss of intelligibility can be measured, in at least this one problem, by non-topological error, but it is quite possible to imagine systems in which non-topological error

would not imply complete loss of intelligibility. Even in radar, such a situation is not inconceivable. It really depends on the purpose for which range information is required and this is a question of meaning. In fact the whole question of defining a natural message dimensionality is one of meaning. So in suggesting that theree is some connection between intelligibility and non-topological error, which itself hinges on the specification of a natural message dimensionality, we may perhaps be merely replacing one vague term by another.

I must conclude by thanking my colleague, Mr. I.L. Davies, for his part in the work I have summarized (6). Between the present version and the more extended one (loc.cit.), there are certain inconsistencies of notation, deliberately introduced here for simplicity.

REFERENCES

1. Shannon C.E. "Communication in the Presence of Noise", Proc. I.R.E., 37, 10 (1949)

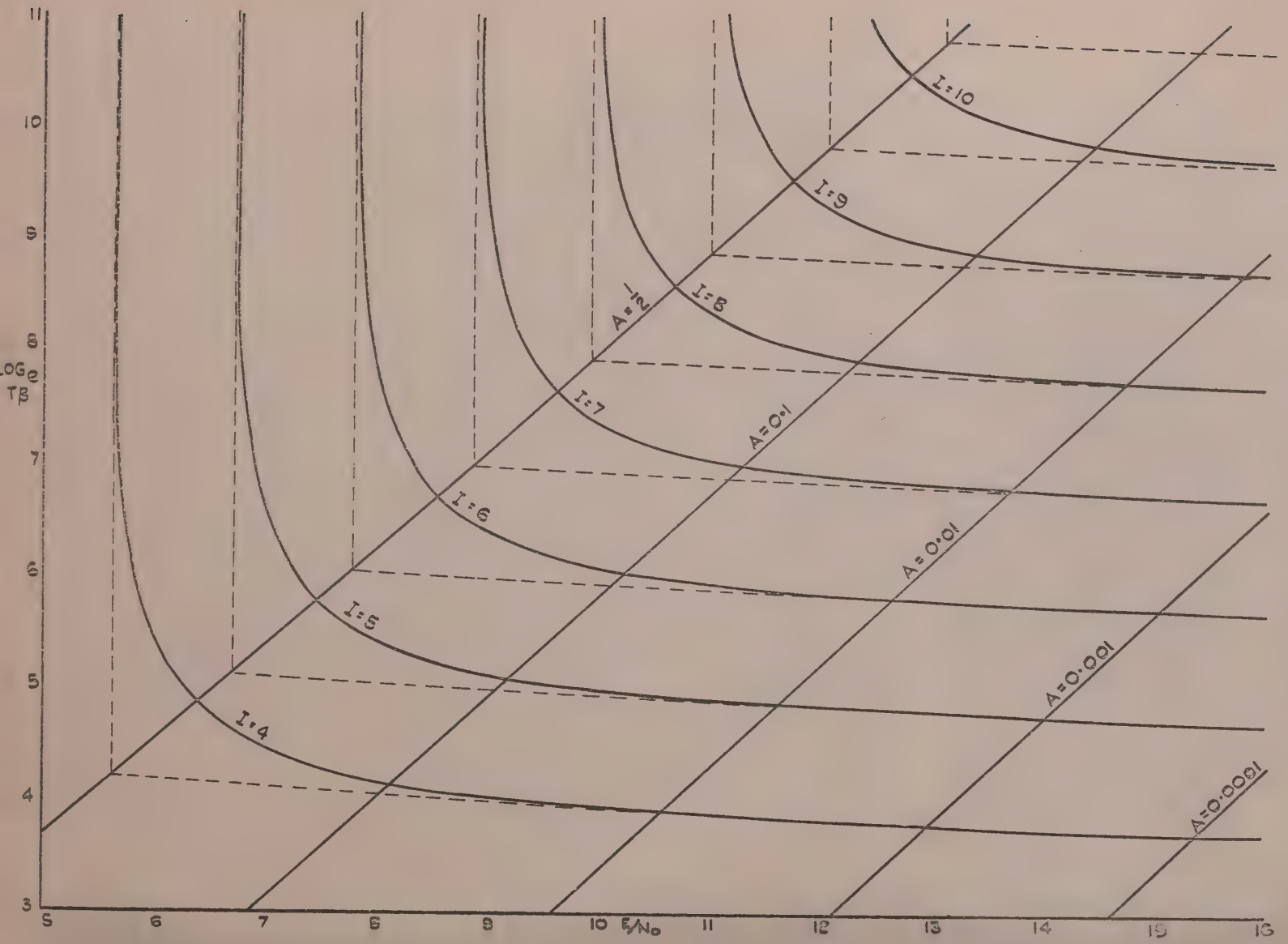
2. Shannon C.E. "A Mathematical Theory of Communication", Bell Sys.Tech.Jour., 27, 379 and 623 (1948)

3. Cherry E.C. Historical Introduction (see page 22)

4. Van Vleck J.H. & Middleton D. "A Theoretical Comparison of the Visual, Aural, and Meter Reception of Pulsed Signals in the Presence of Noise", Jour.App.Phys. 17, 940 (1946)

5. Gabor D. "Theory of Communication", J.I.E.E., 93 (Pt.III), 429 (1946)

6. Woodward P.M. & Davies I.L. "Theory of Radar Information", Phil. Mag., Ser.7, 41 1001 (1950)



INFORMATION AND AMBIGUITY CONTOURS.

T = A PRIORI INTERVAL OF TIME-DELAY

β = WAVEFORM BANDWIDTH (SPECIALLY DEFINED)

E = TOTAL RECEIVED SIGNAL ENERGY.

N_0 = NOISE POWER PER UNIT BANDWIDTH.

I = QUANTITY OF INFORMATION IN NATURAL UNITS.

A = AMBIGUITY BETWEEN SIGNAL AND NOISE.

FLUCTUATIONS AND THEORY OF NOISE

by
D. K. C. MacDonald

(1) INTRODUCTION

What is noise, analytically speaking? Noise - from the point of view of this symposium at any rate - is perhaps best defined as that which may not be predicted with complete certainty. Such a definition may well seem too general to some in view of the fact that in Wiener's theory of prediction (1) one must first remove any perfectly regular (and therefore perfectly predictable) component present in the input; consequently one might conclude that only noise can be left. In the general sense, however, this conclusion is perfectly valid since no physically observed phenomenon can be ideally extrapolated into the future without limit. Thus, for example, if we say that the voltage to be found across the output terminals of a signal generator is:

$$V(t) = V_0 \sin(\omega_0 t + \alpha) \qquad (1)$$

then strictly such a bald mathematical statement implies that throughout the realm of all time, t, this relation provides precisely the value of the voltage. Physically this is of course impossible. If, say, at time $t = 0$ we find the generator to have precisely the frequency $f_0 = \frac{\omega_0}{2\pi}$ and the phase α in concordance with (1)* it will be impossible to guarantee in general that after the lapse of some ten years perhaps these conditions will still be exactly satisfied. To some extent, then, a degree of "randomness" will always exist in the observation of any physical quantity. It is true of course that in many experiments the dependence of the observed variable, y, on various uncontrolled (or uncontrollable) parameters (e.g. temperature, vibration, humidity, gravitational perturbations) is to very slight that during our work we may regard a dogmatic equation: $y = f(t)$: as adequate. Under such circumstances, however, prediction, extrapolation, etc., become mere trivialities and hardly worthy of the names. It is when a degree of uncertainty is present that the "spice" of predictional analysis enters in and all statements in fact which we make about noise in a system are predictions within certain limits. It seems clear in this way why it was inevitable that a statistical framework had to be incorporated into the science of information theory for the present remarkable progress and unification to be achieved. Thus noise analysis may well regard itself as one of the progenitors of prediction theory; at any rate, it bears an intimate relationship to probability theory and statistical mechanics to both of which also information theory claims kinship.

(2) SPECIFICATION OF A FLUCTUATING VARIABLE

If we have an idealised regular function given at all time t by

$$y = f(t) \qquad (2)$$

then from (2) we can determine its rate of variation, for example, at any instant:

$$\dot{y} = \frac{df}{dt} \qquad (3)$$

or the difference between two values of the function at any two points of time:
t_2, t_1: where $t_2 = t_1 + \tau$ say:

---

* Neglecting here any discussion of the inherent compatibility of these two precise statements.

$$\Delta\tau \equiv y_2 - y_1 = f(t_1 + \tau) - f(t_1) \qquad (4)$$

and so on. If also the function is naturally periodic in behaviour, - or may for our purposes be regarded as such -, within a period T, then through elementary Fourier analysis we can obtain a further expression, namely:

$$y = \sum_{n=-\infty}^{n=+\infty}{}' a_n e^{\frac{int}{T}} \qquad (5)$$

where the set of coefficients, a_n, appears to yield further information about y. It is however, important (and often overlooked) to appreciate that (2), (or (5)), contains intrinsically a complete and exhaustive statement of the properties of the function and that further analytic operations yield no essentially fresh data about y, being therefore in that sense trivial. As soon as we permit an element of randomness to enter, then this "triviality" vanishes, essentially of course because (2) ceases by its very nature as a deterministic statement to be a possible representation of the variate. In general, any fresh statement (or "equation") that we make about y, such as its expected magnitude, or average rate of change, will now add significantly to our knowledge of its behaviour, perhaps better to say that we must have acquired knowledge in order to make the statement. In particular, however, it is equally most important to realise that certain statements may be entirely equivalent to one another in their content of knowledge - closely similar to the precise equivalence of (2) and (5) above - although one form may be much more suitable for a particular analytical calculation than another.

(3) THE ELECTRICAL "BROWNIAN MOVEMENT" AS AN EXAMPLE

Let us consider, for preciseness, the case of the electrical Brownian movement ("thermal noise") of a simple linear circuit as sketched in Fig. 1.

The resistance, R, is the idealised localisation of the interaction of the electrical "fluid" with heat, or thermal vibration, characterised by an absolute temperature, T. We wish to inquire about the behaviour of the observable voltage V(t), which, from our knowledge of the random nature of heat interaction, will be expected to fluctuate "randomly". First, from the equiparition theorem of thermodynamics as applied by Einstein to similar problems (2) we may predict the mean square value:

$$\left.\begin{aligned} \tfrac{1}{2}\,C\overline{V^2} &= \tfrac{1}{2}\,kT \\ \text{or}\qquad \overline{V^2} &= \frac{kT}{C} \end{aligned}\right\} \qquad (6)$$

This tells us that if we faithfully observe V(t) for a sufficiently long time ¤ then the mean square value should approximate ever closer to this value. It tells us nothing, however, about its rate of change, for example, and consequently nothing about what we will in fact observe if our measuring instrument is not quite faithful in "following" V(t). Nor does it say anything about the distribution in magnitude of V(t), and so on.

¤ Or, alternatively, average over a sufficiently large number of independent similar circuits.

To gain further insight let us examine the mechanism more closely. C is a pure condenser and must by our a priori laws obey the relation

$$q = VC \tag{7}$$

We are tempted, for R to so say that it simply obeys Ohm's Law

$$\dot{q} = -\frac{V}{R} \tag{8a}$$

but this only expresses the average behaviour of the electrical-thermal interaction. Let us, therefore, (essentially following Langevin(3)) write

$$\dot{q} = -\frac{V}{R} + \vartheta(t) \tag{8b}$$

where $\vartheta(t)$ represents the fluctuating component of the movement of electric charge engendered by the randomness of the heat vibrations. Thus, from (7) and (8b):

$$C\frac{dV}{dt} + \frac{V}{R} = \vartheta(t) \tag{9}$$

We might now follow Langevin further and predict - or assess statistically - the variation of V and $\frac{dV}{dt}$ by "solving" (9) with certain reasonable assumptions about $\vartheta(t)$. Let us, instead, adopt a somewhat different approach leading directly to the concept of auto-correlation. We multiply through by $V(t-\tau)$ (where $\tau > 0$), and take average values (with respect to t).

$$\text{Therefore,} \quad C\,\overline{\dot{V}(t)V(t-\tau)} + \frac{1}{R}\overline{V(t)V(t-\tau)} = \overline{\vartheta(t)V(t-\tau)} \tag{10}$$

Now the auto-correlation function, $\psi(\tau)$, $\equiv \overline{y(t)y(t+\tau)} = \overline{y(t)y(t-\tau)}$ of any "statistically stationary" variable contains a very considerable amount of information about its behaviour. One sees quickly that it characterises quantitatively the idea that a stochastic variable is more -, or less -, random in relation to the time between two successive observations. If these observations are close together (in relation to some characteristic time for the system), then clearly $\psi(\tau)$ will have a significant definite value, while if the time τ is so long that the two observations become uncorrelated - i.e. quite random with respect to one another - then $\psi(\tau)$ will fall to -, and remain thereafter at -, zero. In fact we find that a knowledge of $\psi(\tau)$ is perhaps the closest, - and most directly analogous -, expression for a fluctuating variable to (2) in the case of an ideally regular variable. It is perhaps therefore not surprising to find that a direct equivalence exists also with a frequency relation (the "power spectrum") analogous to (5) (the Wiener-Khintchine Law). Furthermore, successive differentiation of ψ with respect to τ yields information about the behaviour of the time-derivatives of the random variable (cf. analogously (3) following (2)).

Returning to (10), we note that $\vartheta(t)$ is "highly" random representing as it does a phenomenon on the molecular time-scale. Consequently $\overline{\vartheta(t).\vartheta(t+\tau)} \approx 0$ if τ differs significantly from zero, and since V(t) is the physical consequence of $\vartheta(t)$, $\overline{\vartheta(t).V(t-\tau)} = 0$ for $\tau > 0$, and we have now:

$$\frac{\partial\psi}{\partial\tau} + \frac{1}{CR}\psi = 0 \tag{11}$$

Hence

$$\psi = \psi(0)e^{-\frac{\tau}{RC}} \tag{12a}$$

From (6), however, $\psi(0) \equiv \overline{V^2(t)} = \frac{kT}{C}$, thus: (12b)

$$\psi = \frac{kT}{C} e^{-\frac{\tau}{RC}} \quad (\text{for } \tau > 0)$$

and we can readily show also:

$$\psi = \frac{kT}{C} e^{\frac{\tau}{RC}} \quad (\text{for } \tau < 0) \tag{12c}$$

Correspondingly we then have:

$$\phi(\tau) = \overline{V(t).V(t+\tau)} = \frac{2kT}{CR} e^{-\frac{\tau}{RC}} \quad (\text{for } \tau > 0). \tag{13a}$$

$$= 0 \ (\text{for } \tau < 0)$$

$\psi(\tau)$ and $\phi(\tau)$ are sketched in Fig. 2. (13b)

We now have sufficient data to answer our earlier question of what we shall observe if our measuring instrument is imperfect in following V(t). By the Wiener-Khintchine theorem we can derive directly from (12b/c) that the fluctuation is distributed in frequency ("power spectrum") according to the law:

$$\overline{V_f^2} = \dot{w}(f)df = \frac{4RkT}{1 + (2\pi fCR)^2} df \tag{14}$$

Hence if our instrument has a (power) response function: G(f), say then the observed fluctuation will be

$$\overline{V^2_{obs}} = \int_0^\infty \frac{4RkTG(f)df}{1 + (1 + (2\pi fCR)^2} \tag{14a}$$

We cannot yet, however, state how V is distributed in magnitude, - information which would be required, for instance, if we wished to know whether, and for how long, V would exceed some given value. In this case we now appeal to our physical knowledge that the phenomenon is due to a very large number of contributory electron events. This enables us to say that the distribution must be Gaussian ("Central Limit Theorem") and from our present knowledge of the mean square value given by (6) we can then write:

$$p(V)dV = \sqrt{\frac{C}{2\pi kT}} e^{-\frac{CV^2}{2kT}} dV \tag{15}$$

With extended analysis, but making no additional assumptions, we can derive (cf. Rice (4)) the joint dependent probability that $V_1(t)$ shall lie in the range V_1; $V_1 + dV_1$ and $V_2(t + \tau)$ shall lie in V_2; $V_2 + dV_2$, namely, the two-dimensional Gaussian distribution:

$$p(V_1, V_2, \tau) = \frac{C}{2\pi kT(1 - e^{-2\tau/RC})V_2} \exp - \frac{C(V_1^2 + V_2^2) - 2V_1V_2e^{-\tau/RC}}{2kT\ (1 - e^{-2\tau/RC})} \tag{16}$$

Thus, to summarise this investigation, equations (6), (12), (15) represent states of progress in our information about the properties of V(t) and enable us to predict more comprehensively its behaviour, while on the other hand (12) and (14) are entirely equivalent statements of the same property suited to two different analytical situations.

(4) THE FUNDAMENTAL SIGNIFICANCE OF NOISE

In many situations to-day noise itself - as its name suggests - is regarded as an essentially undesirable intrusion whose influence on our measurements we wish to minimise. In this case, analysis such as the foregoing may enable us to design our apparatus in such a way that this end is in fact achieved. On the other hand, it seems wise here to emphasise that a fluctuation itself may be a powerful source of information. The classic example is of course the original Brownian movement which finally settled the molecular theory controversy engendered by Ostwald's school of "Energetics". Again to-day the observation of electrical noise in crystal rectifiers, for example, can provide valuable aids to the theoretical investigations. In particular the "brain-waves" observed by the electro-encephalographer could well at first have been called noise, and we may interpret the regularities observed in the auto-correlation function of the records as appropriate to the transmission system of the brain whereby they are communicated to us. Another excellent case is that of radio-signal reflection from the ionosphere. For many years the prime purpose of transmitting a signal upwards and observing the returned wave was the determination of the effective height of the reflecting layer; irregular "fading" of amplitude and variation of phase were essentially just nuisances both in the research field and in conventional broadcast reception. The recent recognition (5) that this "fading" had much in common with electrical noise as observed through a circuit of limited frequency-response has led to an appreciation of the variable, or stochastic, element in the make-up of the ionosphere. By then translating the established analysis of random noise into the appropriate concepts one can deduce, for example, the root mean square velocity of the scattering elements from detailed observation of the fading-rate of the returned signal. Again there has often been controversy over the validity of the classical equations for the simple shot-effect - that is, the electrical noise induced in a "saturated" (temperature-limited) diode valve as a direct consequence of the granular nature of electricity. In particular, the validity of a determination of the quantum of electric charge from measurements has been questioned in the past. It appears in fact much more fruitful to regard again such work as yielding valuable information about the physical conditions governing the emission of electrons from a cathode surface; in this case close agreement with the simple formula

$$\overline{\delta I_f^2} = 2eIdf \qquad (17)$$

indicates that, in fact, electrons are emitted as individuals independently from one another.

The general situation is in truth typical of physical research. When first the wave nature of X-rays was in question as a hypothesis it was suggested that a crystalline lattice ought to act under suitable conditions as a diffraction grating; from the approximately known inter-atomic spacing and the diffraction pattern observed (if any) it ought then to be possible to deduce an estimate for the wave-length of the X-rays. Once, however, the hypothesis was established beyond question the situation was reversed and X-rays were used as a tool of immense value for deriving precise information about the lattice structure of suitable diffracting solids.

(5) ENTROPY IN FLUCTUATION THEORY AND INFORMATION THEORY

Since information theory may logically be regarded as a specialised branch of statistical-mechanics or -mathematics, we may well expect to find many conceptual links. In particular, the concept of statistical entropy has been applied to information theory and this causes us to raise here the question of the status of entropy in a system subject to fluctuations. It appears that not a great deal has been done in this respect-essentially because the entropy will be effectively constant in a stationary statistical system - although Shannon has defined formally a noise entropy and shown, for instance, that Gaussian noise has maximum entropy for a given mean square deviation. However the case, for example, of a long electron beam in a modern transit-time valve presents a problem of particular interest. It can be designed so that the beam will enter the "drift-space" more or less fully ordered in density; then, as it travels, it must ultimately exhibit the shot effect, due to the intrinsic disordering influence of the thermal velocities (6). Concomitantly any regular signal initially imposed on the beam will finally be dissipated by the same phenomenon. Qualitatively we may certainly say that the entropy has increased but the precise analytic relationship for this parameter has not yet been worked out*. A satisfactory expression might enable much of the available apparatus of statistical mechanics to be brought to bear directly on such problems particularly where forces of interaction have to be reckoned with. We should note, however, that a full specification of the variation of Γ^2 or "space-charge reduction factor", embraces a more detailed knowledge of the fluctuational behaviour of the beam than would be given by a knowledge of the entropy alone.

REFERENCES

(1) Wiener, N. "Extrapolation, Interpolation and Smoothing of Stationary time Series" p.150, Wiley, New York (1949).

(2) Einstein, A. Zeits. fur Elektrochem. 13, 41, (1907).

(3) Langevin, P. C.R.Acad.Sci., Paris, 146, 530, (1908).

(4) Rice, S.O. Bell Syst.Tech.Jour., 23, 282, (1944); 24, 46 (1945) - Secns. 3.1 and 3.2 particularly.

(5) Ratcliffe, J.A. Nature, London 162, 9, (1948).

(6) MacDonald, D.K.C., Phil. Mag., 40, 561, 1949.
Phil. Mag., 41 863, 1950.

* I am grateful, however, to Dr. C. Domb, Oxford, for discussions in this field.

FIG. 1.

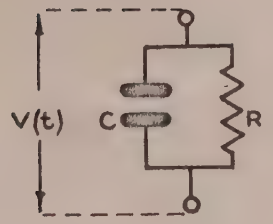
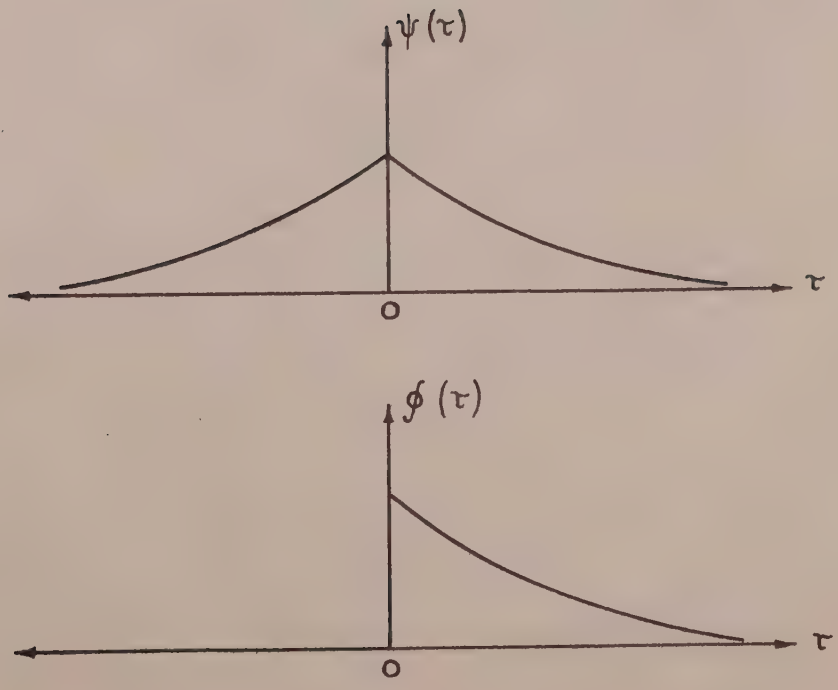


FIG. 2.



COMMUNICATION THEORY AND LINGUISTIC THEORY

by
D.B. Fry

In the previous papers and discussions in this symposium there have been many references to language and to the properties of languages and the similarity between the subject of communication theory and the subject matter of linguistic theories is already evident. This short paper will attempt to set forth, in a very summary fashion, some of the aspects of linguistic theory which might prove of interest to those who are primarily concerned with communication theory.

Any language, whether it be naturally evolved or deliberately constructed for some purpose, is an obvious example of a symbolism or 'representation' and this fact inevitably forms the starting point of most linguistic theories. In the last years of last century, Ferdinand de Saussure, in his Cours de linguistique générale, enunciated the principle of the dual nature of what he called the 'linguistic sign' and employed the terms 'signifiant' for the symbol and 'signifié' for the concept which is symbolised. It was inevitable that, both before and after his day, a great deal of time and energy (and no small amount of printers' ink) should be expended in the study of the relations between the 'signifiant' and the 'signifié'. Now the difficulty of knowing what one is talking about, which is inherent in the discussion of any subject, is considerably aggravated in the case of linguistics and semantics, since language is the only medium in which to discuss language. In particular, the study of semantics seems to lead inescapably at some stage to 'the meaning of the meaning of meaning', the 'semantics of semantics' or some such infinite regress. It is fortunate, therefore, that by common consent the discussion of communication theory is not concerned with the semantics of the messages of signals which are under consideration: we are concerned only with the code or codes by which the meaning is conveyed.

The purpose of this paper is twofold: it is on the one hand, to examine some of the characteristics of the codes in an existing language and on the other, to make some speculations about the means which may be employed in the individual listener for decoding the messages which come in. The examples used will be drawn mainly from English and it is important to note that we shall be concerned with the auditory form of English, not with the visual, written form. When it is necessary to represent the auditory language visually, a phonetic transcription will be used, that is a system of letters having a one-to-one correlation with the units of the auditory language.

Probably the easiest way in which to approach the whole question is to visualize as far as we can the chain of events when two people communicate with each other by speech. At the opposite ends of the chain we have the mental activity of the speaker and the listener, which we may call for convenience, though rather inaccurately, conceptual thinking. The ultimate aim of the speech communication is that the conceptual thinking of the listener should proceed along lines which are at least similar to those of the speaker. Intervening between the thinking of the speaker and of the listener we have a communication chain which is composed of a large number of links, and at each link we have a transformation which constitutes an operation of communication. The whole process may be divided into three main sections: the encoding of the message in the speaker, the transmission of the coded message through the medium of communication, and the decoding of the message by the listener. The second of these will not concern us at present; we shall consider only the operations which take place in the speaker and the listener and of these, only those which take place at the higher levels of the central nervous system, since other speakers are dealing with the working of the receiving apparatus nearer to the periphery.

It is evident that at present any discussion of what happens at these levels in the speaker and the listener must be almost entirely a matter of speculation. This paper is concerned only with showing how the existing body of linguistic theory may form a basis for such speculations. The continuous

signal which enters the human ear as the physical input, gives rise eventually to quantised information in such a form that it can be assimilated into conceptual thinking. The input is, of course, quantised in one particular fashion at the stage of transformation into nerve impulses along the acoustic nerve. The linguistic analysis of a language suggests that at higher levels than this we should imagine a series of transformations of the incoming signals into units of different magnitude, and suggests that the operation of decoding in the listener may be considered as a succession of scanning processes in series, the output of each scanning providing the input of the next, and the successive outputs having the form of increasingly larger linguistic units.

The first of these stages with which we are concerned receives as its input the sensation patterns set up by the nerve impulses to the brain. These sensations seem to have four dimensions: pitch, loudness, quality, and length. These terms refer only to sensations and not to any feature of the physical input to the ear. It is for this reason that 'length' is used here to denote the psychological correlative of 'duration' in the physical world. These four terms then denote the basic dimensions of acoustic sensation and the input to the first scanning stage may be visualised as patterns of sensation bounded by these four dimensions. It is interesting to note that in the language of the world we find examples of the use of each one of these four to carry a sematic difference: that is to say there are pairs of semantic units (or concepts) which are represented by pairs of sensation patterns differing from each other with respect to one of these dimensions only. In all languages the differentiation of such pairs by means of quality is common (as for example in the English words put, pat) but we find also, for example, in Finnish, differentiation by length only; in tone languages such as the Chinese dialects differentiation by pitch, and occasionally, in Serbo-Croat, differentiation by loudness.

The input of the first scanning stage is therefore a series of four-dimensional sensation patterns and at this stage the patterns are compared with those already existing in the brain and corresponding to the smallest linguistic units, the phonemes. These are the units of the auditory language which correspond approximately with the letters of the visual language. Thus a segment of a message corresponding to the word 'thinks' would yield at this stage a series of phonemes θ, i, ŋ, k, s. The groups of phonemes form the input to the second stage and here the output symbols are the morphemes of the language. The morpheme is the unit which fulfils what we ordinarily understand as a grammatical function and thus a message such as: 'He thinks he's waiting too long' would yield the following morphemes:

/hi/ /θ iŋk/ /s/ /hi:/ /z/ /weit/ /iŋ/ /tu:/ /lɔŋ/

In the next stage the output symbols are simply the words of the language in the everyday meaning of the term and the message which we have just decoded at the morpheme stage would give the six words which make up the message.

Finally the groups of words are identified as sentences, which form the output of the last scanning stage, and in this form the message is assimilated to the conceptual thinking of the listener who is then aware of the 'meaning' of the message.

Each of these scanning operations may be considered as involving a communication channel; we have in each case a set of input symbols, a transformation and a set of output symbols. We are therefore concerned in effect with a series of codes for each language and we shall now examine the kind of knowledge about the codes which can be supplied from the results of linguistic analysis. This analysis attempts to establish 1) the levels of analysis appropriate to the particular language (since

all languages are not necessarily susceptible of analysis at the same number of levels), 2) the individual units which the language employs at each level (we may also in addition have some knowledge of the frequency with which given units occur in the language), 3) the factors which govern the choice of a specific unit in a given sequence of units. We can translate this into terms of communication theory and say that we may discover the number of codes used, the units in each code, possibly something about the probabilities associated with any given unit and lastly, a good deal about the constraints which operate on the units at each level. In order to exemplify these points we shall consider some of the available information about the auditory language in English.

We have seen that we have to deal in English with four distinct codes. Of these, it is the phoneme level which has been studied more extensively than any other, and it will thus provide the majority of the illustrations in this section. The number of units in this particular code is about 40 for most types of English pronunciation. If all phonemes were equi-probable in any position in a sequence there would be a probability of 1/40 that any one would occur, but we can get nearer to the true probability if we know the frequency with which every phoneme occurs in a large sample of English speech.(1) The frequencies of occurrence of phonemes in a very common type of English are known and this gives us the a priori probabilities for the phonemes, but without regard to their position in the sequence.

It is obvious, however, that in a language we are concerned with sequences in which the probabilities at each choice are dependent upon the choices already made and our table of frequency of occurrence gives us only an approximation to the probability of a phoneme's occurring at a given point in a sequence. In order to make a more accurate calculation of the probability we should need to know a good deal about the constraints which operate upon the symbols in the code. Here again linguistic analysis provides a certain amount of information. For example, in English, if we allow the sentence pause as one of the signals in the code, then the probability that a given phoneme will occur next to the pause is not that given by the frequency table since at least one sound (ŋ) cannot occur in this position at all. Similarly in the position before a sentence-pause, the sounds h, j, w and r (in the pronunciation we are considering) have zero probability, as have also a number of vowels, e, a, o, u, . Sequences of consonants also reduce the probabilities at succeeding positions, so that after a pause, groups pr, gr, spl, str are possible but not sequences such as pd, kp, slp, trs etc.

This kind of information could be gathered also at the other levels of analysis. For example the morpheme /iŋ/ in 'waiting' could be replaced in other sequences by a certain limited number of other morphemes /s/, /id/ and /zero/, each with a certain probability. When we go from the phoneme to the morpheme level, we encounter a very great increase in the total number of symbols in the code, since the number of morphemes is approximately equal to the number of words, which may for an average person be of the order of 100,000.

The step from words to sentences introduces another great increase in the number of signals as the total number of possible sentences, even for one single speaker, is very large indeed. But again, at each level, the uncertainty as to the occurrence of any unit is immensely reduced by the operation of constraints. One of the most obvious of these is what we may term the 'line of thought' of the speaker; this scarcely needs any illustration since it operates widely in all our communications with each other. 'I have just seen a man with a red beard'. Here is a sentence which had a very low probability indeed in the present sequence because it was virtually ruled out by the line of thought which was being pursued.

(1) Fry, D.B., "The Frequency of Occurrence of Speech Sounds in Southern English", Arch. nēerland.d. Phon.Exp., 1947, XX, 103.

It is interesting to see how constraints operating on one level entail similar effects at other levels. The theoretically large number of sentences which is possible at any moment is in practice reduced, not only by the line of thought but also by what J.R. Firth has called the "context of situation". This factor works at the sentence level and also at the word and phoneme level. We most of us learn fairly in life that there are certain things which cannot be said "in the drawing-room". This is an illustration of the context of situation operating to reduce the number of possible sentences and words. But further than this, the situation also has an effect on the phonemes which may occur since particular pronunciations are appropriate to particular situations. One might, for instance, imagine a political speaker addressing a public meeting and saying: "If you ask me, ladies and gentlemen, what the present government has done to ease the lot of the rodent operative, I can only reply 'Dunno'." Now here the sequence dənou is an impossibility because of the situation, although the sentence is a perfectly possible one.

Another very common example of the influence of situation is to be found in the adjustment of the listener to the particular pronunciation of the speaker. When an Englishman listens, for example, to an American speaker, the 'context of situation' causes him to insert a different repertory of symbols into his scanning machine at the phoneme level.

The effect, then, of the context of situation and the line of thought is that the theoretically very large number of sentences which are possible at any moment is very much reduced and that this constraint on the choice of sentences is reflected in the choices at the lower levels - i.e. at the levels where the linguistic units are of smaller magnitude.

This fact leads us back to a consideration of the scanning operations by which, we have suggested, the process of decoding might be carried out in the listener. The knowledge of the situation and the line of thought reaches the listener mainly by way of the lower scanning levels and is stored at the highest level, so that we have to visualize a system of feed-back circuits by which this information has its effect on the functioning of the lower levels. This is true, not only between the highest and the lowest levels, but also between each level and the next below, since this is the only way in which the system of constraints in the successive codes could be implemented. The output of a given scanning operation is used to set up the required repertory of symbols to be scanned in the succeeding operation. To take an example at the phoneme level, if the series of output symbols were h, i, z, s, k, this information - or as much as is necessary - would be used to determine that the set of possibilities for the next phoneme would not include p, t, k, b, d, g, etc., etc., but would include for example most of the vowels and r.

Such a system is equally necessary to explain the self-correction which operates in the listener's decoding machine. If at some stage an error in decoding occurs, this fact usually appears at a subsequent stage and the correct symbol is inserted in place of the incorrect one. For example, a sequence of words: 'Can I jump my books here' would be corrected by feed-back from the sentence level and the word 'dump' substituted for 'jump'.

This general picture of the decoding process, although highly speculative, does not involve impossibilities, at least in principle, and we may hope eventually to gather something in the nature of evidence about the way in which the brain carries out the decoding operations. One source of such evidence which is already being investigated is the behaviour of aphasic patients in whom it is possible to study the effect upon speech and understanding of brain lesions. The picture is inevitably complicated in these cases because it is generally true that both the reception and the transmission of speech is affected, but one can none the less gain some indications that the brain does deal with the linguistic symbols at different levels. In some patients, for example, one finds transpositions at the

phonemic level; in other cases, the patient may be incapable of dealing with any unit smaller than a word; or again, a defective child may be capable of reproducing a whole repertory of speech sounds (equivalent to the phonemes) and yet be unable to form them into sequences which make words and sentences.

In conclusion, a few remarks concerning the redundancy of auditory English may be of interest. By adopting the technique used by Dr. Shannon it is possible to gain some light on this matter at the phoneme level. Experiments in which a subject was asked to guess the next sound at each position in a connected sequence have shown that the redundancy of the auditory language is less than that of the visual; the number of sounds guessed correctly at the first guess was about 55 p.c. as compared with the 75 p.c. obtained with a written text. This difference is to be expected and is explained by the anomalies of English (and even American) spelling. One of the most interesting results obtained was the clear demonstration that the redundancy on the phoneme level is least at the word boundaries and inevitably increases in proportion to the length of the word.

(At the end of the paper a demonstration of redundancy in auditory English was given, in the form of a sound recording of the following texts. In the first, only one vowel quality was used throughout but the distinctions between consonants were preserved; in the second, one consonant only was used, but vowel distinctions were preserved.

(1) ð- s-mz t- b- n- d-t ð-t -ŋgl- -z kw-t -nd-st-nd-bl

-f -nl- w-n v-l -z j-zd pr-v-d-d ð- k-ns-n-nts - r-t-nd.

(2) o- -i ʌ -o -a- i- -i -u:- ou--i -ʌ- o-ɔ-ə-

-i -ai- ə -ʌ- -ei-ə -o- ə- i-e-i-i-i-i-i.)

HEARING

by

T. Gold

Sense organs perform generally two different functions; in order to make the distinction between them clear, we may think of a technological "sense organ", like a television camera. The information contained in the incident light is there received by a great number of channels, namely the elements of the mosaic. But the output is required to be conveyed by a single channel. Accordingly the requirement is not only a change of the "physical variable" - light to electricity - but also such a treatment upon the information as to adapt it to the transmission mechanism which is to follow.

In the case of the television camera, as in most technological devices it is preferable to transmit information through one or a small number of channels, although, of course, the bandwidth in each channel will be inversely proportional to their number.

In biology the state of affairs is different. Multiplicity is common, but the information handling capacity of any one channel, such as a nerve fibre, is strictly limited. It is hence not surprising to find organs whose function of adapting the information to the transmission mechanism consists of splitting it up and distributing it to a large number of channels.

In the case of the organ of hearing the requirement for such a functio is obvious. Sound, represented by the fluctuations of air pressure, is presented to the ear along a single channel. The ear has then to transmit the information contained in these fluctuations to the brain. But the range of fluctuation speeds is some thousand times faster than that which can be handled by a nerve fibre. Some process of splitting up and distributing the information to a large number of channels (i.e. nerve fibres) must hence occur. Of the large variety of possible processes it is the analysis into frequency components which the ear employs. Sufficiently many bands of staggered frequency are used to cover the entire range, and each such band is sufficiently narrow for a nerve fibre to be able to cope with its information content.

A further analysis according to amplitude also takes place; for again a nerve fibre cannot signal as great a range of amplitudes as the ear is confronted with. Several nerve fibres of staggered amplitude response are therefore used for each of the frequency bands.

If one wishes to investigate the mechanism quantitatively one can rely to some extent on subjective experiments. For although one has no knowledg of the loss of information which may occur in the brain, it is certain that the ear must be competent to transmit at least that information of which we do become aware. Such experiments can be used to give a good indication of the bandwidth, or the reciprocal quantity which is the oscillatory time-constant, which must be associated with the elements of the frequency analy

At first one would think that the measurement of the smallest interval of frequency which can be recognised would constitute such an experiment; but there are severe difficulties in the interpretation. For if the two slightly different notes are sounded simultaneously then the phenomenon of beats provides an accurate clue which is independent of the mechanism; and they are sounded in succession then it can still be argued that an accurate comparison of the amplitudes of many elements of the analyser may enable a change of frequency to be discovered which is only a small fraction of the (conventionally defined) bandwidth of any one element. One may compare thi with an out-of-focus photograph of a black line, where accurate photometry of the plate would enable the true position of the line to be reconstructed The precision to which amplitudes must be known for such a reconstruction t be effective would, however, soon become absurdly great if the apparatus we tested with many closely adjacent signals simultaneously.

Professor Pumphrey and I have carried out such an experiment. A waveform of sound was accurately produced, whose spectrum contained a large number of closely spaced maxima and minima, fitting under a wide envelope curve. A second waveform differed from this in that its spectrum had the positions of maxima and minima interchanged, but without change of the envelope curve. The experiment now consisted of finding the closest spacing of the maxima and minima for which a distinction could still be made between the two signals.

The two waveforms possessing these properties consisted of sinewaves regularly interrupted by short silent intervals; and one was so constructed that the phase was continuous from one pulse to the next, whilst in the other a phasechange of 180° was introduced in each silent interval. Now it is clear that the interpretation of the experiment can be described in two ways, which are essentially identical. We can either say that we found how small the bandwidth of the analyser elements had to be in order that there should still be a significant effect of the individual peaks of the spectrum; or we can say that we found the interval of time that has to elapse between two oscillatory signals before the relative phase of the two becomes irrelevant. Pursuing the second description, we derive a lower limit to an oscillatory time-constant (and hence an upper limit to a bandwidth). We may regard the experiment as revealing "phase memory", that is showing for how long some part of the mechanism continues to oscillate, after the impressed signal has ceased.

We found in this way a lower limit of oscillatory time constants of the order of 10 milliseconds. Now it is important to point out that the experiment does not bear any other interpretation, and that it was possible to check that no other unwanted effect enabled the distinction between the signals to be made.

Similar values of oscillatory time-constants or bandwidths would be deduced from a large number of other experiments, if they were interpreted very simply; in particular from the experiments of Galambos and Davis on single fibres of the auditory nerve. But one has been hesitant to adopt this interpretation, for it seemed to lead to a conflict with the physical interpretation of the cochlea. The degree of damping of the elements of a mechanical frequency analyser constructed on such a small scale, and filled with liquid, can easily be seen to be much greater than would be compatible with such narrow bandwidths. And it seemed proven that the actual frequency analysis is carried out mechanically.

The way out of the dilemma, it seems to me, is not to require that the mechanism shall be explicable as a passive device. A local supply of energy could provide each of the elements with the necessary amount of positive feedback to counteract the damping to any desired extent; and in fact some prominent by-products of the action of the cochlea show the existence of such a local energy supply, and fit in well with the suggested feedback. It is not permissible to test this suggestion, as would be convenient, on a dead cochlea, for it is known that the subsidiary effects which must be linked with the feedback action have ceased there.

The structure of the cochlea is not too complicated to allow calculations to be made of its mechanical properties, taking into account the possibility that the individual elements may be effectively only slightly damped. The analyser elements are in series with respect to the sonic signals, and it can be shown that this leads to an asymmetry of the frequency response curve of each element, such that it exhibits a sharp cut-off towards the high frequencies, but only a much gentler slope towards the low frequency side. This is in agreement with all that is known in that respect; many subjective experiments show such an asymmetry, but it is particularly clearly demonstrated in the objective measurements of Galambos and Davis.

REFERENCES

1. Galambos, R. & Davis, H. "The Response of Single Auditory Nerve Fibres to Acoustic Stimulation." J. Neurophysiol. 6, 39. (1943)

2. Gold, T. & Pumphrey, R.J. "The Cochlea as a Frequency Analyser." Proc. Roy. Soc.(B) 135, 462. (1948)

3. Gold, T. "The Physical Basis of the Action of the Cochlea" Proc. Roy. Soc.(B) 135, 492. (1948)

condition, [illegible]
[illegible] nerve to the central nervous system. In every case [illegible]
[illegible]sion is in a succession of electric pulses propogated [illegible]
fibres, and the only independent variables are (a) the partic[illegible]
involved (b) the time sequence of the pattern of pulses in it.

When the sense organ is required only to signal a simple local condition e.g. to indicate how much a muscle is stretched, this information may be transmitted rather exactly. But when the organ is concerned with the whole visible world, only a minute fraction of the total information is transmitted. The question "what is coded and how" is one which is worth the attention of the Meeting.

The vertebrate eye is essentially a photographic camera in which an inverted real image is projected upon the retina. It is received upon a mosaic of 120 million rods and 6.5 million cones (in man). If each of these photo-receptors were connected to a separate optic nerve fibre it might be possible for the brain to receive information as detailed as the retinal mosaic. But since there are only a million optic fibres, this is far from being the case. Only the cones of the very central part of the retina have 'private lines' to the interior of the brain, and even this applies only to man, monkeys and birds. The rods are connected to a single nerve fibre by clusters of some hundreds, and though this arrangement would not baffle a communications-engineer, it is certain that the single nerve fibre does not transmit a message from which the contributions of all the separate rods can be reconstructed. Our inability to appreciate detail when looking at the periphery of the visual field or upon a dark night (when rods only are active), confirms the conclusion that detail present at the photoreceptors has been lost.

The interesting question is "In what way is it lost?" There would seem little point in having so fine a mosaic of receptors if all to be extracted was the summed or average response of them. Nor does the nerve structure of the retina support this. For the convergence of hundreds of rods onto one optic fibre does not occur simply. There is such a tangle of intertwining threads and cross connections that details have never been clearly seen. The retina is only $\frac{1}{4}$ m.m. thick and in such small dimensions, there is opportunity for nerve interaction by the diffusion of electricity or chemicals as well as by the propagated pulses. Whatever be the mechanism, it is certain that light falling upon one part of the retina will affect the nervous response from light at another part. Thus within the retina itself there is an elaborate mechanism for interaction between signals from one photoreceptor and another - an interaction which has taken place before the impulses reach the optic nerve where the great reduction in the number of lines occurs. It seems reasonable to suggest that the interaction may be the encoding of certain aspects of the detailed image received by the rod mosaic. What aspects?

This is not easy to answer by introspection because introspective analysis can be satisfactorily applied only to processes taking place at the conscious level, which retinal processes are not. Nearly the whole of the visual region of the cerebral cortex in man is concerned with the little 'private line' part of the retina, and no doubt this fact is related to our conscious impression that the rest of the retina doesn't much matter.

But though our question is hard to answer through introspection, it might yield rather well to the right kind of experiment. This can best be

done upon animals (vertebrates) because the electric code in a single optic nerve may be directly recorded in response to various kinds of visual stimuli. It is not difficult in fact to collect a bewildering array of such coded messages. In order to interpret them we need to find what invariant in the retinal event corresponds to a fixed code. It is only practicable to do this by having in mind a number of ideas as to what abstractions of the retinal event are likely.

Here then is the problem, put rather loosely (lest by defining too closely one excludes some of the very considerations which may help towards the answer).

The outside events are projected in space and time upon the retina which itself is seldom at rest. The retinal receptors can respond individually with good resolution and interact with one another to encode certain features of the situation. This message is sent along a single line supplying some hundreds of receptors and subject to the usual limitations of nerve transmission e.g. no frequency above 500/sec.: no variation in amplitude.

What are the features of the situation which, with these limitations, it would be most advantageous to abstract and transmit - advantageous in the sense that it will allow the animal not only to make automatic movements; but also to form the basis of visual judgement - the squirrel jumping in the trees, the dog catching tossed food - for these animals have no 'private line' system.

One would guess "movement" to be the prime visual abstraction - movement which may mean food or may mean foe. Even stationary objects are turned into movement by a flick of the eye. Can Theory substantiate this kind of guess-work and refine it and lay down the conditions in which experimental observations will give definite answers?

INFORMATION THEORY IN PSYCHOLOGY

by
W.E. Hick.

Although the title of this paper suggests something like a general survey, its real theme is a particular example of Information Theory applied to Experimental Psychology. Since such applications have otherwise been practically non-existent so far, it may be that one concrete example will convey more than the equivalent amount of general speculation.

The problem is that of explaining Reaction Time, especially the so-called Choice Reaction Time. Briefly, reaction time is the minimum time it takes a given person to respond, by some simple voluntary movement, to a pre-arranged signal. More exactly, the situation involves instructions to the subject, his presumed comprehension of them and cooperation, and a warning signal, or its equivalent, preceding the operative signal. The time is measured from the beginning of the stimulus or signal to the beginning of the response.

This simple experiment was invented in 1850 by no less a person than Helmholtz, so it has now reached its century. It has been modified and elaborated in all sorts of ways, reasonable and unreasonable, but it has managed to preserve its individuality - and its mystery; because nobody has explained why it is what it is - why it is so little affected by changes in the nature of the stimulus and response, for instance, and why it is so long.

Many reflexes, involving only lower levels of the nervous system than are required for voluntary action, act far more quickly, although one cannot say that they are any simpler, in terms of stimulus and response. But, of course, the reflexes are more specific - they work in their own particular way, to their own particular stimulus. Whereas at the top level - the level of voluntary movements - we can use innumerable different kinds of signal or response, and we get substantially the same result. All that matters is that the stimulus shall have a clear and abrupt beginning, and that the response shall be something the subject can already do at will - something already part of his repertoire.

One of the early elaborations of this simple reaction was the Choice Reaction. Here the subject is instructed that one out of a particular set of different signals will be given, and he has to make the correct response, according to a pre-arranged system. He does not know which of the set of signals will be given on each occasion, but they may occur with equal or with different frequencies, according to the purpose of the experiment. The number in the set will be called the Degree of Choice.

As long ago as 1885, reaction times were measured for all degrees of choice up to 10. The subject waited with his ten fingers each on a Morse key, and was to press the appropriate key when a symbol was suddenly exposed. The results have been plotted in Fig. 1, together with those for repetition of the same experiment performed by the writer. As one might expect, the reaction time increases with the degree of choice. Each point plotted is, of course, the mean of a large number of readings. No previous attempt seems to have been made to fit a function, but it can be seen that the points lie remarkably near a smooth curve which, apart from a difference in scale, is the same for both experiments. Moreover, nobody seems to have tried seriously to account for the relation.

It is not yet profitable to frame a hypothesis in strictly physiological terms. But, bearing in mind that the brain, like a computing machine, consists of a vast number of nearly identical units, there is perhaps a prima facie case for asking whether any operation which it performs within itself - its response to a terminated disturbance, so to speak - may usefully be regarded as a sequence of elementary

the less need is there to take account, in its construction, of the probability distribution of possible signals. With respect to the discrimination between "signal" and "no signal", the brain is very reliable indeed; a failure rate of one in 10,000 is a highly conservative estimate. So it is not quite outrageous to ignore the probability question here; whether doing so amounts to assuming equiprobability depends on the point of view adopted.

If we represent the supposed process in the form of the well-known "tree" diagram (Fig. 2), we are reminded of a difficulty in picturing what happens as a sequence of narrowing classifications. As long as we only think of complete stages - that is, if the tree, whatever its size, is always complete and symmetrical - the logarithmic relation obviously holds. But if the effective number N + 1 does not happen to be an integral power of 2, the matter is not quite so simple. If N + 1 is 6, say, the next complete tree has a capacity of 8, and if that is used, all the signals will have to go through 3 stages of analysis. Some of them will be over-analysed, in fact, and the reaction time, on this view, will be the same as for N + 1 = 8.

The result would be that the reaction time would go up in steps, of equal height but exponentially-increasing length. There is no sign of this in the data, although it could hardly fail to show, especially in readings from only one subject, where it would not be obscured by individual differences. We could, however, imagine a tree with some of its terminal twigs lopped off, as in (b), Fig. 2. Then, if N + 1 is 6, 4 of the signals would go through 3 stages, and 2 would get only a 2-stage analysis. If the signals were given with equal frequencies on the whole, as in these experiments, the result would be an average number of stages slightly greater than $\log_2 6$. The curve, in fact, would touch the logarithmic curve where N + 1 was a power of 2, but elsewhere would bulge above it (Fig. 3). The maximum discrepancy, however, would be about 0.086 of a discrimination-step, regardless of the size of the tree, and it would need an excessively large sample of readings to show it. In any case, it may not be present at all; if the tree is, so to say, flexible, so that signals can potentially be routed by the shorter path, it requires only a small degree of prediction on the part of the subject to eliminate the bulges. Adequate randomisation of sequences can cope with this in theory, but again it demands an enormous number of readings.

It is interesting to note, in passing, that if the tree is distorted to the limit of asymmetry, it can represent the operation of systematic searching, as shown in (c), Fig. 2. There is one long branch and a number of side-branches. The extreme asymmetry accords with the low efficiency of this method, as noted above.

It must be emphasised that the tree diagram is not intended to represent the neural mechanism involved. As far as present knowledge goes, it can be best be only a picture of the average process. Two factors - the large variability and the subjective impression that the reaction times depend almost entirely on attention and expectancy - hint that the simple tree diagram may not, by itself, be a satisfactory representation of the process. But it is an interim representation, where before there was not even that.

It may be added that practice is necessary to the point of being able to make the right response without using a conscious formula. To achieve this for the higher degrees of choice necessitates many hundreds of reactions, and is incredible tedious. And it would be quite impossible for very high degrees; for instance, to respond correctly to one of a row of 1,000 lights could only be done by counting, and the logarithmic law would give place to something nearer a linear law. With regard to practice and the formation of automatic responses, it can be seen in Fig. 1 that the 1885 data for the 9 and 10 choice reactions fall below the logarithmic curve. It is not known whether an unduly large number of mistakes were committed in these cases [illegible]

operations of like kind and like duration. Such a shot in the dark can be justified, if at all, only in the sequel. The answer will depend mainly upon whether the definition forced on the expression "of like kind" happens to fit easily into some relevant field of knowledge. The field of Information Theory is probably relevant to the work of a mechanism constituted as the brain is, and therefore it is to be hoped that the phrase in question can be given a simple "information-definition".

Thus, in the case of the choice reaction, the essential process is one of recognition or identification, which is analogous to matching the signal with one of a set of standards or gauges. For example, the correct gauge might be found by a form of searching, of which there appear to be two principal methods, namely (a) the purely random, which gives an exponential distribution of numbers of trials required, and (b) the systematic, in which no gauge is tried more than once, and which gives a rectangular distribution. But the average number of trials is proportional in both cases to the number of gauges in the set. Each trial is an operation of like kind, in the information sense, and might be expected to take the same time. Consequently, on this analogy the reaction time should be a linear function of the degree of choice, which is not the case.

The search method has also the teleological disadvantage of being inefficient. It extracts too much information, in the sense of determining a number of particular things which the signal is <u>not</u>, as well as what it <u>is</u>. In other words, the trials are binary discriminations which do <u>not</u> have the probability of 0.5 associated with them.

In the choice reaction time experiment one can, of course, evaluate at once the amount of information required to be extracted, assuming that successful prediction by the subject is negligible. In the experiments cited, the different signals were given with equal frequencies, and therefore the amount of information is some logarithmic function of the degree of choice. If the hypothetical elementary operations are extractions of units of information, reaction time should then be logarithmically related to the degree of choice (N). Clearly, however, the relation cannot be simply

$$RT = K.\log.N$$

because that would give zero time for the simple reaction ($N = 1$). We may add a constant equal to the simple reaction time, but it is difficult to see any theoretical justification for doing this. (There should, of course, be a <u>small</u> additive constant in any case, to allow for such things as transmission time to and from the brain, but this is likely to be less than 1/10 of the simple reaction time, and is not considered here.)

However, there is another point to be borne in mind, and this is that although instructions and training do, to some extent, enable the brain to concentrate on the set of n signals, it does not withdraw entirely from awareness of all the other extraneous information continually arriving at the sense organs. The analytical mechanism must, therefore, first determine that the experimental signal is one of the pre-arranged set, and not something irrelevant. This is obvious in the case of the simple reaction, for, although there is only one signal in the set, it must convey information, and it can only do so if there is at least one alternative; in fact, there is one general alternative - that whatever is happening at the moment is not the signal awaited. And similarly, for the higher degrees of choice, there are logically $N + 1$ relevant possibilities. Consequently, although it is a long shot, we have some justification for expecting reaction time to be proportional to $\log. (N + 1)$, and, as Fig. 1 shows, that is approximately true.

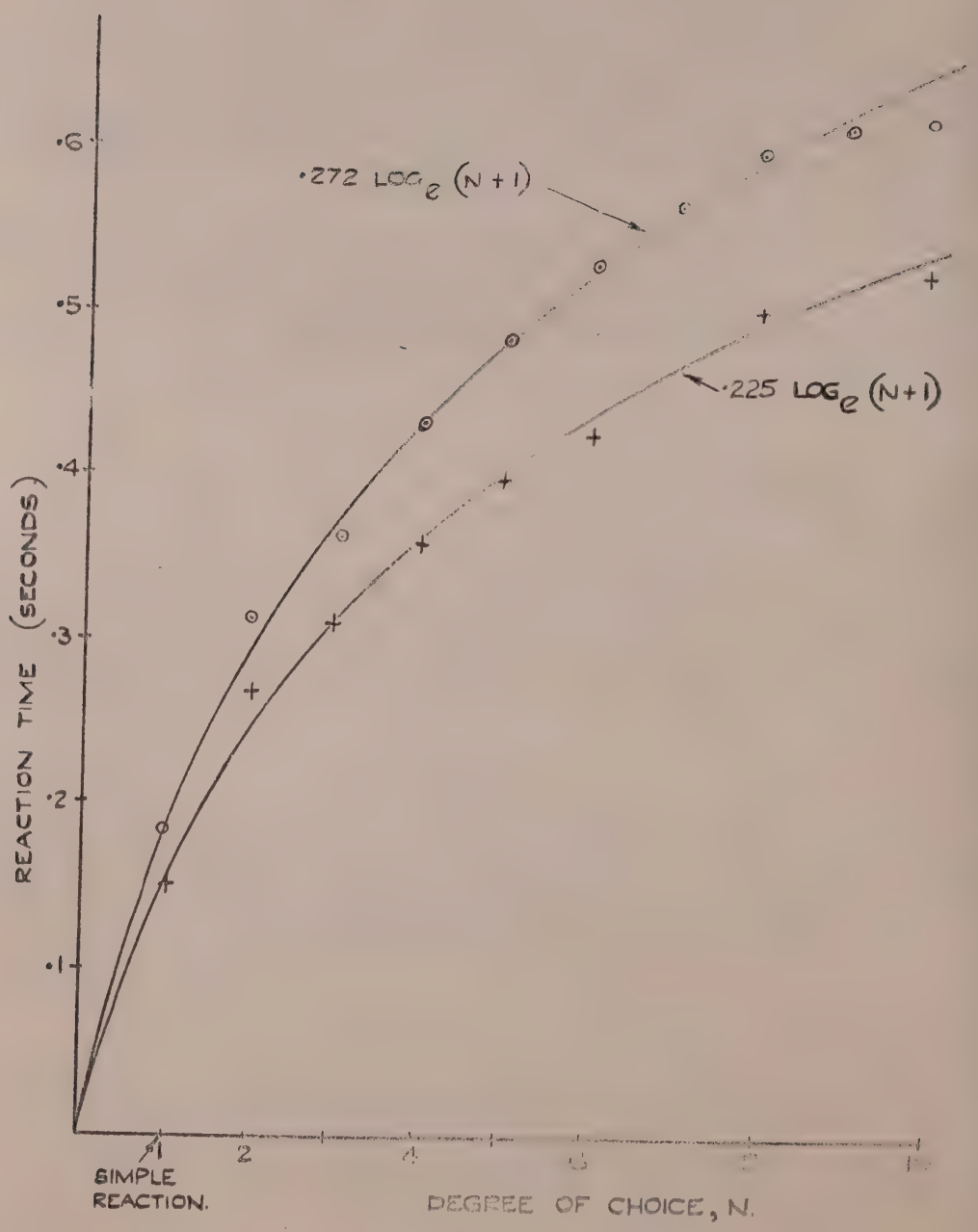
It will be noticed that nothing has been said about the relative probabilities of the n signals and the one extra. Of course, nothing <u>can</u> be said, except that the more reliable a discriminating mechanism is,

is quite likely, and would imply that some signals were not being fully analysed. If they were one stage short, they would still have a 50 per cent chance of releasing the correct response, and consequently a proportion of improperly short reaction times would be recorded. In the case of the writer's data, there was no such increase of mistakes, and the divergence from the curve is within the chance expectation.

Finally, with regard to the future, what has been done so far is only a beginning. For instance, it is hoped to continue with different types of control and display code. Also, in the experiments so far, the effect of the signal was to reduce entropy suddenly from a maximum to nearly zero. By different techniques, different degrees of entropy change may be obtained, and this may throw further light on the matter. Apart from this, it is interesting that the tree concept, which in some sense is equivalent to coding in a binary number, helps to explain why an item of information closely following a previous item is apparently held up for some fraction of a second until, so to speak the analyser is cleared.

It is useless to speculate on the outcome of these projected developments, or on more general possibilities, but it is hoped that what has been said will suggest at least some of the ways in which Information Theory may be brought into Experimental Psychology.

FIG. 1.

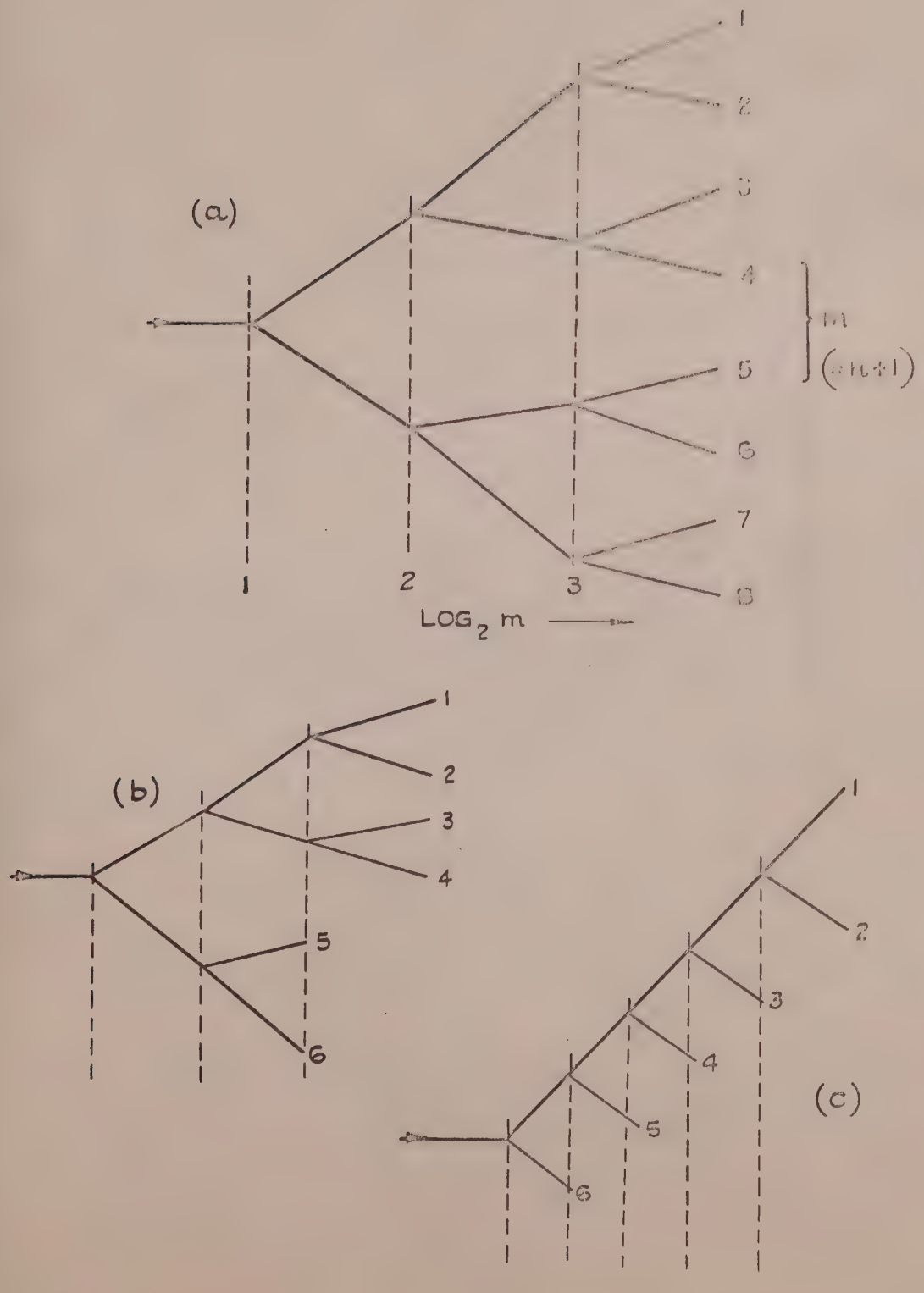


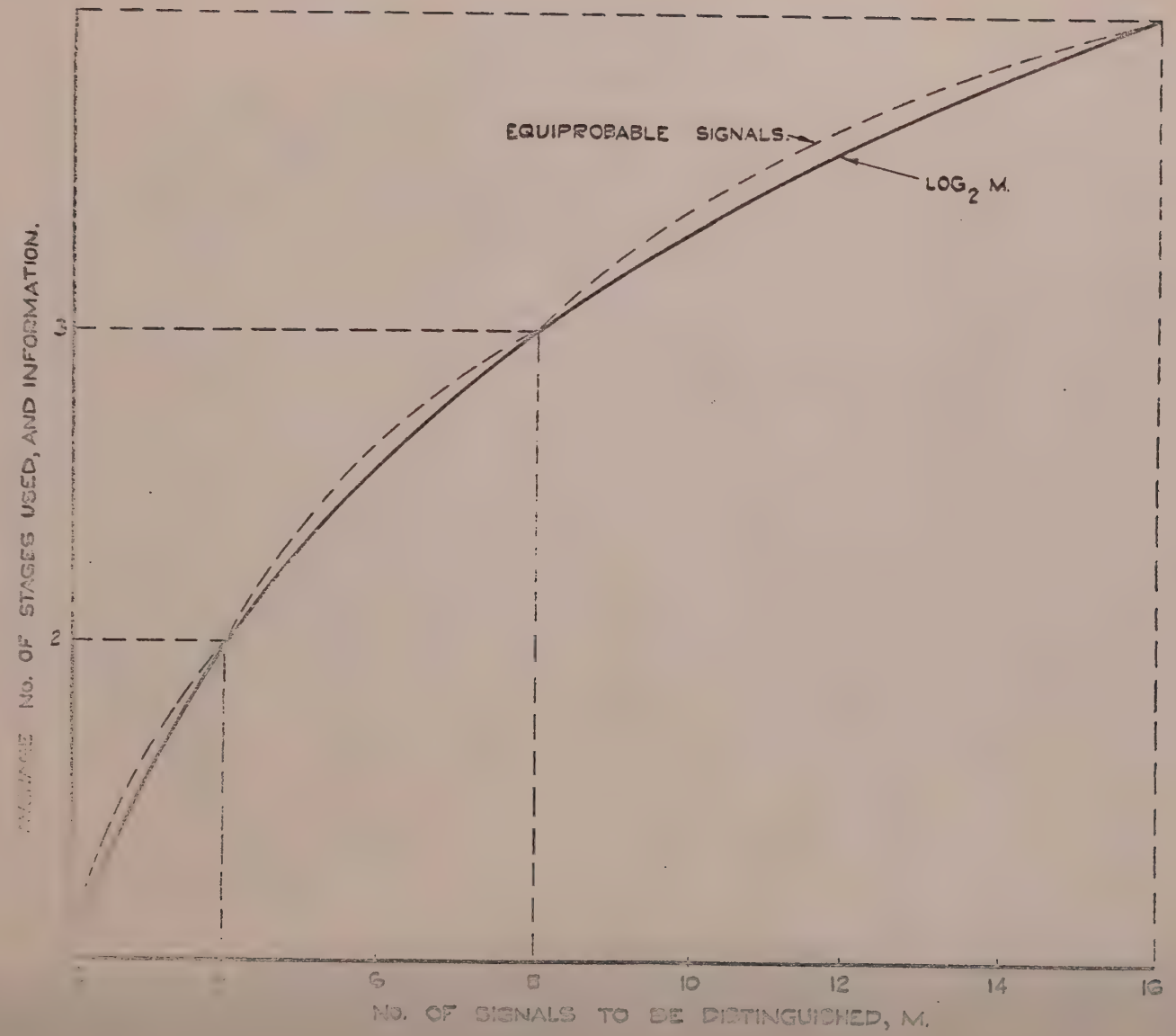
CHOICE REACTION TIME

⊙ - MERKEL (1885, 9 SUB[illegible])

\+ - HICK (1950, 1 SUBJ[illegible])

FIG. 2.




EQUIPROBABLE SIGNALS.
LOG_2 M.
AVERAGE No. OF STAGES USED, AND INFORMATION.
2
6
8
10
12
14
16
No. OF SIGNALS TO BE DISTINGUISHED, M.

POSSIBLE FEATURES OF BRAIN FUNCTION AND THEIR IMITATION

by
W. Grey Walter.

The electrical activity of the brain has been studied intensively for about 20 years. Psychological and biochemical studies have a larger history still, but nevertheless there is very little knowledge of what the brain does in physiological terms. The success of electro-encephalography (EEG) in clinical problems is due mainly to empirical correlation with proven disease-states, very little to inference from experimental data. One may say that what the human brain does is everything we can personally experience or do, but this does not assist understanding. One might as well say that the number and affinities of the chemical elements must be such as to account for all known substances. A discouraging feature of the human brain is that not only does it have a stupendous number of functional modes; it also contains about 10^{10} nerve cells, many of which may be capable of any sort of connection with any or all of the others. If this be the true scale of the problem it were vain to seek any but a statistical answer.

There is, however, some reason, to believe that the functional element is much larger and the number of such elements much smaller than would be indicated by cell counts. A possible number is of the order of 1000. These elements may be arranged so that, in a system of two, A and B, the following modes can exist:

$$(0),\ (A),\ (B),\ (A + B),\ (A \rightarrow B),\ (B \rightarrow A),\ (A \rightleftharpoons B).$$

For large numbers of elements the number of possible modes is given approximately by $M = 2^{(n^2 - n)}$, where n = number of elements. Thus in a brain with about 1000 elements the number of modes would be about 10 raised to the three hundred thousandth power - a sufficiently large number to account for our subjective impression of thought complexity and variety.

It is also helpful to consider what functions a mechanism such as a brain might be expected to have, if it be considered as a device for the construction of compact, plastic electrochemical models of the universe in which it finds itself. The structural information content and coding of the models is obviously dependent in the first instance upon the limitations of the sense organs and such instruments as may be available to assist them. Their metrical information is likely to be projected on a variety of co-ordinates so as to permit ready parameter transposition and transformation; data from one sense organ must be comparable with those from another, particularly if the brain is to compose and store linguistic representations.

An important principle which emerges at an early stage of such speculations is that of parsimony; as applied to animals this means that organic evolution tends rigorously to eliminate redundant organs or functions. Similarly a single element or organ tends to participate in a number of functions. At the same time a factor of safety is maintained such that about half of an organ or function can be destroyed without causing immediate death.

With these ideas in mind, models have been constructed using the smallest number of elements which can provide an imitation of the superficial features of animal behaviour. With two receptors (transducers), two nerve elements (relays), and two effectors (motors), behaviour patterns appear which are so complex that in practice the future of the system cannot be predicted accurately from external knowledge of its past. Slight uncertainties are cumulatively amplified by the richness of interconnection of the elements. These models display the following behavioural properties which will be described and illustrated.

1. Searching (scanning). In the absence of an adequate stimulus (signal) such a light, the device is in constant movement and scans its horizon until a signal is received or its power supply is exhausted.

2. Positive Tropism (Vectorial feedback). When an adequate light signal is received the scanning process is halted and the steering servo directs the model toward the light source.

3. Search for Optima. When the intensity of a light source is greater than a certain level, the steering servo is again brought into operation so that the model avoids the source and circulates round it.

4. Avoidance of Buridan's dilemma. When two equal light sources are equidistant from the model, its scanning and optimopetal mechanisms ensure that it will approach first one and then the other, and ceteris paribus will oscillate between the two.

5. Negative Tropism. When a material obstacle is encountered a circuit is formed which changes the internal amplifier into a multivibrator, leading to alternate butting and withdrawal movements combined with a change of direction, so that the obstacle is either displaced, surmounted, or circumvented. The obstacle is "remembered" for about one second.

6. Discrimination. While behaviour mode 5 is in operation all other modes are impossible, so that the model is indifferent to distant positive stimuli while in contact with immediate negative ones.

7. Internal homeostasis. When the internal power source is nearly exhausted mode 3 is abolished so that the model can approach close to a light source, if suitable connections are available the power store can then be recharged; during this process all other modes are impossible.

8. Self-recognition. A pilot light source is connected in the scanning-steering servo circuit. If the model receives a signal from its own light source reflected in a mirror, this source is itself extinguished; the signal is thus abolished and the scanning-steering servo is again connected but this restores the signal and so forth. Oscillations occur which generate a specific pattern of behaviour.

9. Mutual recognition. Two such models, receiving signals from one another's pilot light, each extinguish their own, and again complex oscillations arise between the two models generating a characterustic "social" behaviour pattern.

LEARNING BY ASSOCIATION

The addition of two or three more valves confers on the device the power to establish and extinguish simple conditioned reflexes. The minimum requirements for this process may be enumerated as follows:

1. Differentiation with respect to time of specific signals.
2. Extension with respect to time of neutral signals.
3. Mixing of coincident specific and neutral stimuli.
4. Summation in time of coincidences.
5. Activation of preservation system by summed coincidences.
6. Preservation with slow decay in time of information that a certain number of coincidences have been observed.
7. Combination of preserved information with fresh neutral signal to form new response.

The electronic circuits performing these operations are quite simple; the behaviour of models with them parallels very closely that of simple animals in the conditioned reflex laboratory. Neuroses appear and subside slowly with rest, more quickly with shock, permanently when the learning circuit is removed altogether. The need for three separate time parameters for this apparently simple function is noteworthy.

In this model, the preservation process is mediated by a damped oscillation with a slow decrement; if not reinforced it subsides to below threshold in a few minutes. After each reinforcement it is restored to its original level. After a certain number of reinforcements it can become permanent.

Experiments now in progress suggest that the structures and mechanisms responsible for these seven operations in animal brains may be identified and isolated. The first four operations are essentially components in the statistical appraisal of the environment; the last three constitute memory and recall.

SIGNIFICANCE OF INFORMATION THEORY TO NEUROPHYSIOLOGY

by

J.A.V. Bates

Lashley (7) has recently reminded us that Descartes not only had the concepts of structural modification of living tissues to account for memory, but also clearly had before him the concepts of the scanning mechanism. It is as well therefore to begin by considering in what respects our ideas are in advance of those of Descartes. Only in one respect I believe are they essentially in advance of his. We now have as a background to our thoughts the concept of evolution, and together with it the notion that certain attributes of living matter have survival value. The phenomena comprised by the term "mind" are included among these attributes, with the result that, as Professor Adrian (1) has recently remarked, we have a "tendency nowadays to regard the relation between mind and matter as one which need not give rise to much difficulty". The idea, inconceivable to Descartes, now commonplace among biologists that the phenomenon of mind is a byproduct of matter in action dates, as we see from the historical introduction to this symposium by Cherry (4), at least to Julien de la Mettrie (1740). In addition to this, present-day biologists have a kind of modesty which would be quite foreign to Descartes, for we realise now that our own picture of the universe is extremely partisan. This modesty is forced on us not only by the realisation of the fact that our sense organs have a severely restricted sensitivity to the various ways in which energy can manifest itself, but also we realise that the types of things we can think about and the ways in which we can think about them, are determined by the layout of our nervous system. For example, as a member of the primates, we have a disproportionately large amount of our forebrains devoted to vision, and we are denied the kind of interpretation of nature that would be made if this was replaced by the sense of smell. Nevertheless, in spite of these advances, many of Descartes' difficulties are still with us. The problems of the mechanism of memory and of pattern recognition or 'stimulus equivalence' are today a particular stumbling block. And Information Theory is I believe particularly relevant to Neurophysiology because some of our difficulties may exist through an inadequate understanding of the problems of coding, and of making representations; problems which appear to be the 'bread and butter' of "Information Theory".

We are all here agreed that it is in every sense desirable that electronic engineers should speculate in neurophysiological problems, and by way of encouragement I will illustrate briefly some of the pitfalls which may beset speculation in this field. The first of these is failing to observe that if a generalisation is put in a sufficiently tight form, it only needs the demonstration of one clear exception to upset it. For example, as Lorenz (8) has recently written in exactly this connection, "if J.B. Watson had only once reared a young bird in isolation, he would never have asserted that all complicated behaviour patterns were conditioned". Secondly, there easily arises confusion through the unfortunate choice of terminology; for example, in the use of the word redundency in Information Theory. On closer inspection we realise that redundency is a rigidly defined and measurable attribute of the ensemble from which a message is selected, but the word carries with it an emotive tag and implies something undesirable in the ensemble which may be quite erroneous. Or again, one might point to the use of linear system terminology and particularly that based on sine-wave technique, to cover obviously non-linear mechanisms - for example the use of such terms as phase-advance and frequency modulation in contexts where they obviously cannot strictly apply. Thirdly, it is well to recall how similar terminology may cloak differences, for example both electricians and neurophysiologists will use the word pulse or impulse to describe signals which differ in their rates of propagation by a factor of 10 to 100 million. Lastly, do not forget that there are other models besides electronic ones which may be helpful in biological problems. For example, some students of animal behaviour are concerned with the origin and

control of "nervous energy" or "action specific energy" in their experimental animals. They find it convenient to picture this commodity as obeying the physical properties of fluids, a pastime which Lashley has dubbed "psychohydraulics". But Lashley's wit is apt to veil a genuine difficulty in accounting for some aspects of reflex and instinctive behaviour. Do not, therefore, be surprised if you run against biologists who are inclined to pay respect to Schroedinger's (12) view that the physical properties on which a piece of living matter operates may well be as remote from those of current electrical theory, as the physical principles on which a dynamo works are remote from those of the theory of heat engines.

Neurophysiologists will be interested in your speculations according to the extent to which they suggest experiments, and so it is as well to have some notion of the difficulties which surround experimentation in this field. I would say that there are three levels; firstly, at a purely technical level - I need not enlarge on this since the difficulties here will be obvious to you when you consider firstly that the essential units are nerve fibres of the order of 100th of a mm. in diameter or less, and that although the electrical changes associated with an impulse may be in millivolts, the nerves are surrounded in the body by tissue fluids which act as an electrical shunt; and secondly, that the normal function of nervous tissue is extremely easily disturbed by interference for example with its blood supply, and large and important parts of the brain are practically inaccessible to electrical recording without disturbing this activity.

At the second level we have what might be called methodological difficulties, for example we have the difficulty of interpreting data derived from experiments involving the behaviour of animals. This is well illustrated by a recent story of Lashley's (7) of an experiment in which some monkeys were trained in a variety of visual discriminations which they had to signal by pulling on appropriate strings. They were then operated upon and it was found that extensive destructions in different areas of fore-brain all caused loss of these habits. They were then re-trained in a task involving the discrimination of weights, and as soon as they had learned this new task, the habits based on visual discriminations returned spontaneously. I am not prepared to say exactly what the interpretation of this experiment should be, but its significance to my point is that when the monkeys failed to show their old visual discrimination habits after the operation, it might well have been concluded that the operation had destroyed some "association pathways", if the subsequent experiment had not been performed. This story too, will illustrate two other pitfalls. In the first place it is said that the monkeys were re-trained in the discrimination of weights it would have been more correct to have said the experimenter considered that monkeys were discriminating weights, because differences in weight were the most obvious differences to his own sensory organs. Nevertheless, it is an inference to say that the monkeys were in fact discriminating weights, we cannot be certain without more evidence that they were not, for example, discriminating differences in smell or using some other sensation which escaped the experimenter because it was to him sub-threshold. Secondly this story can be taken to illustrate the pitfalls which may beset unwary delving into the literature in search of data to support a hypothesis. It is as well to remember that the interpretation of any literature which is not strictly in our own field is a very risky business, particularly when that field is in general less rigorous than the one we are familiar with.

The third class of difficulty in neurophysiological research is difficulties of concept. I will illustrate this by outlining two problems where besides other difficulties, conceptual difficulties seem prominent. The first of these concerns what has been called the plasticity of the nervous system, which shows itself for example in a special case known as the transfer of learned reactions. To illustrate what I mean I may start at the highest level and ask you to imagine a musician, playing a flute concerto with an orchestra, before a microphone.

We will imagine that he has learned the concerto by ear and has no music. He has played so far every note correctly when suddenly a piece of chewing gum becomes dislodged and blocks his flute. Being a resourceful man he continues to whistle the remaining notes, and the listeners are unaware of the accident. If you will accept this story as not totally inconceivable, you will see that it disguises an attribute of considerable physiological significance. We can safely assume that he has never whistled the notes before, nevertheless he achieves a comparable end result by using a totally new and previously unused set of muscles. What he has clearly been learning during his practice is that a series of musical notes and intervals must follow each other in appropriate sequence, and he must do his best to achieve this irrespective of the actual muscles he uses. In the same way, having learnt to write your name with your hand, you can at least make a fair shot at writing it with your nose. At the other end of the scale perhaps one might instance the phenomenon of a spider repairing its web. Not only does each new web present a new structural problem unlike all webs that have ever preceded it, but each time a web breaks the problems of repair will be different. The spider if observed will be seen to use the total range of its possible muscular movements until a satisfactory repair has been achieved. Likewise we can note the vastly varied movements of an ant endeavouring to hide its egg after its nest has been turned over, or of a newborn kitten seeking its monther's nipple. These examples illustrate what appear to be an extremely widespread attribute of living matter, namely that under certain conditions there is some "non-invarience" in the muscles used to achieve a particular end result. In other words, the conception of a sort of internal pianola roll playing out a pattern on selected muscles is obviously inadequate. I suggest that in order to account for this phenomenon it may be necessary to suppose that within the animal there is present at any instant a physical representation of a particular state of its environment which is required to be signalled by its nervous system at some time in the immediate future. For example, try to hum or imagine yourself humming "God Save The King" at about one beat per second and introspect what is happening. I am suggesting, as a first approximation, that at the instant you are sounding any particular note you are carrying a representation of the sound you wish your ears to be receiving one second in the future. I am not wishing to imply a particular rigidity in the future time scale, but I am suggesting that animals operate to the order of seconds ahead rather than days. Apparent evidence of long term planning, for example nest building, may be due to unwarranted inferencew from our own behaviour. Craik (5) among others considered that we carry within ourselves a physical model of reality, but the model to which he referred was a posterior one, delayed at least by nerve conduction time say 10 m/secs. The essential notion behind the idea of behaviour seems to me to be that the concept carries with it a requirement within the animal for some anterior model of reality, to which a posterior model can be matched.

It is pertinent to observe here that this manifestation of what has been called the plasticity of the nervous system is not invariably found. The most interesting and significant exceptions to it can be shown to exist where in a sense one might expect on utilitarian grounds they should exist. For example, the gray-lag goose, in order to retrieve an egg which has rolled out of the nest, encircles the egg with her neck and lifts it back. Lorenz (8) has shown that when the normal egg is exchanged for an artificial one similar in appearance but lighter in weight, the goose, having got the egg on her neck makes a movement which is too strong and fails to replace the egg in the nest, and she does not appear to be able, by practice, to modify the strength of this movement to achieve success. Similarly if the artificial egg is heavier than the normal egg, the goose is unable to lift it off the ground, although she possesses extremely powerful neck muscles. But it should be noted, that in order to bring out this defect in the design of the goose it requires the intervention of an altogether higher order of intelligence, the product of a primate brain. The goose, like the ant, is adequately

designed to cope with any disturbances to its nest that might reasonably occur in a state of nature. I believe that the idea of revealing design defects by this type of experimentation is a hopeful one for telling us more about the design of man. It has of course to some extent been exploited by Gestalt Psychologists.

It might seem likely that in order to perform a complicated sequence of co-ordinated muscular acts all the bulk of the animal's brain would be needed, but this is not necessary. For example, Bard (3) has shown that the complex behaviour of a female cat mating can be induced in animals with both cerebral cortexes removed, (i.e. in bulk about 80% of their brain), by injections of oestrin, a chemical substance which can be synthesised and which like other hormones produces comparable changes in a wide range of animals of very different structure.

The second problem to which I will call your attention is the problem of memory. I will confine myself to a few brief points. In the first place, before one starts to consider this subject it is necessary to rid oneself of the complications introduced into it by the phenomenon of language. The fact that I can ask you if you remember what you had for dinner yesterday and thereby set your thoughts travelling in a particular direction is clearly a completely trick phenomenon in a biological sense, a by-product of our sophisticated powers of communication, and one which at the moment need not concern us. In the second place, we must recognise at the outset that our terminology is confusing because there is a difference between the state of having a memory and being unable to recall it, and not having a memory. This difference can be clearly demonstrated in man by the use of drugs and hypnosis etc. and I see no reason to doubt its existance in animals. Thirdly, we come to what seems to be the hard core of the subject, namely that animals can form associations between cause and effect, and whatever these associations are they must have some physical representation. Lashley (6) goes so far as to say that all creatures from worms upwards can form this association at a single trial, and he has stressed that animals are graded not by the fact that they can form associations, but by the complexity of the kinds of things which they can associate. Von Bonin (14) has pointed out that this complexity is correlated more closely with the ratio of the weight of their brain to the weight of their body than with various other indices which have been proposed, for example, the volume of the frontal lobes. 'Complexity' of association has not so far been a measurable quantity, it appears to be to some extent synonymous with a spacial or temporal remoteness. It is significant that we have intuitively no scale of orders of complexity, associations when they have been formed all seem equally 'obvious'; this is well illustrated in the story of Hardy the mathematician who, having made some complex mathematical deduction, remarked that "it is obvious'. He then retired for half an hour and returned saying "yes, it is obvious". It is inconceivable that this process in Hardy was not of a different order of complexity to the associations which he or any of us formed in infancy. The peculiar fact about these associations is that they exist in the animal, dated in time and space, for example if I asked you to call up the memory of Nelson's Monument you will visualise something resembling a rod in the same plane as the long axis of your body. If you were to change your position in space by lying down then you can correct for this and imagine the column at right angles to your long axis. Likewise it is obvious that we have something corresponding to a private time scale to which our memories are fitted. How this comes about is entirely obscure, one can say that any system which behaves in some consistent way with respect to time is in fact a clock of a sort, thus then neurone is a clock since time could be measured by changes in polarisation of its membrane.

It is a curious fact that the evidence at the moment is conflicting as to whether or not these associations can be located at any particular place in the animals brain. Broadly speaking I believe it is true to say that the idea that each part of the cerebral cortex has a separate and

definable function is losing ground, this tendency is illustrated by comparing Figures 1 and 2. Fig.1, taken from a paper by Putnam (11) quoting Kleist (1931), and Fig.2 from Penfield (10) dated 1950. Even though the later one is based on much better evidence, there are difficulties even in accepting this as representing an invariable state of affairs in a normal person. It has been arrived at chiefly by exciting the surface of the brain with an interrupted current in a conscious patient at operation. But there is such a vast amount of inter-connection within the brain that it does not follow that a function is located at an actual point where you can excite or abolish it. On the other hand, a considerable amout of data, both from animals and man, in which there has been extensive destruction of some part of the brain by disease, injury or operation are broadly in agreement that loss of brain tissue so far as it affects behaviour brings about a loss of the ability to generalise, to abstract and to modify behaviour by the consequences of prediction, and in general to regress to a somewhat lower and less highly adaptive form of behaviour. Generalisation in these terms appears to hold true in a way which is dependent rather more on the amount of cortex which is destroyed than on the precise location of it. These findings, together with those of an increasing brain/body weight ratio in higher animals, rather lead one to the belief that what appears to us as a single association may in fact be a representation of a number of discrete items each of which has a separate physical representation in a separate part of the brain. In this way the more complex associations would be represented by a greater number of discrete items requiring a greater bulk of brain, and removal of a particular part of the brain might effect to some slight extent a number of different associations rather than totally abolishing certain of them. But I must confess however, speculation on this subject along these lines does not suggest to me at the moment any new experimental approach, although the parallel with Mackay's (9) use of the concept of the atomic proposition is tempting.

There are however two lines of experiments in this connection which have been suggested by this symposium. In the first place, Uttley (13) has suggested a way in which the theory of information may be used to measure the degree of patterning or structured arrangement in a collection of dots, and this suggests that it might be interesting to see whether the ability of different animals to distinguish patterns was correlated with the degree of structured arrangement of these patterns as measured by the information content.

In the second place, it is now possible to say that the brain receives information from all the sensory surfaces in two forms. These Adrian (2) has called temporal and spacial information, temporal information being the number of impulses per second in a particular fibre, and this is related to the intensity of the stimulus; spacial information being the actual fibre out of a large number of possible fibres which is conducting the impulse. Thus spacial information is derived from an analysis of the stimulus carried out at the sensory surface. In the case of the eye, it is easy to visualise, since a particular fibre corresponds to excitation arising in a particular point in the space around the animal. In the case of the ear, a particular fibre corresponds to a particular frequency (pitch) in the stimulus. In the case of the nose, a particular fibre corresponds to a particular class of stimulating molecule. It is technically possible to prepare and manipulate electrodes of the size of a single fibre and to stimulate nerves at an operation with the patient conscious. It may therefore be possible to get more information about coding in the nervous system by introducing impulses into a fibre in a known code and asking the patient to report his sensations.

To conclude, there are a few other aspects of neurophysiology to which information theory may be relevant. In the first place we have heard that the optic nerve contains 2 million separate fibres.

Let us assume that at the end of each fibre there is a mechanism sensitive to 10 different modes of impulse frequency in that fibre. Since the fibre can conduct impulses up to about 400/sec. this estimate may well be conservative. In this case the total number of separate patterns which could be conveyed by the optic nerve would be the number 2^{10^6}. The utterly astronomical nature of this number may be realised when compared to Eddington's estimate of $2^6 \times 10^5$ for the number of electrons in the universe. It is not unusual that as soon as we commence quantitative consideration we reach astronomical numbers of this kind. It can sometimes be said that the body is constructed on a fairly liberal basis with plenty of spare in case of injury, but this is not always true, particularly in the case of the optic nerve in the above example, where a diminution of 10% in the number of fibres would probably be noticeable. Clearly these vast numbers must have some interpretation. Is it in this case, an unavoidable by-product of the requirement for improved resolving power?

Secondly, it is clear that any general statement that can be made about a measuring apparatus is also, under stated conditions, applicable to the behaviour of animals. For example McKay's (9) generalisation of the concepts of band width and logon capacity to include the behaviour of the microscope makes it now meaningful to talk in terms of band width and logon capacity of the behaviour of an observer whose gaze is directed into the eyepiece. Interpretation on these lines might have relevance to an experimental investigation of visual sensory thresholds in man. Lastly one might observe that birds clearly have the problem of communicating in the presence of noise, the cuckoo for example provides a simple illustration of some aspects of 'redundancy'. Perhaps an examination of birdsong on the basis of information theory might not only suggest new types of field experiment and analysis, but also provide a fresh stimulus to human communication engineers.

REFERENCES

1. Adrian E.D. "Physical Basis of Perception". Oxford (1946) p.92.
2. Adrian E.D. Brit. Med. Bull. 6. 330. (1950)
3. Bard P. Am. J. Physiol. 116. 4. (1936)
4. Cherry E.C. "A History of the Theory of Information" (see page 22).
5. Craik K.J.W. "The Nature of Explanation" Cambridge (1943).
6. Lashley K.S. Quart. Rev.Biol. 24. 28 (1949)
7. Lashley K.S. Soc. Exp. Biol.Symposium No.4.(Cambridge)1950.p.454.
8. Lorenz K.Z. Soc. Exp. Biol.Symposium No.4 (Cambridge)1950.p.233.
9. Mackay D.M. Phil. Mag. Ser.7 41. 289 (1950)
10. Penfield W. quoted by Jefferson G. Brit. Med. Bull. 1950. 6. 333.
11. Putnam T. Ass. Res.Nerv.Ment.Diseases (Res.Pub.No.19) (1939)p.81.
12. Schroedinger E. "What is Life?" Cambridge (1944)
13. Uttley A. "Information, Machines and Brains" (1950) (see page 143)
14. Von Bonin G. J. Gen. Psychol. 16. 379. (1937)

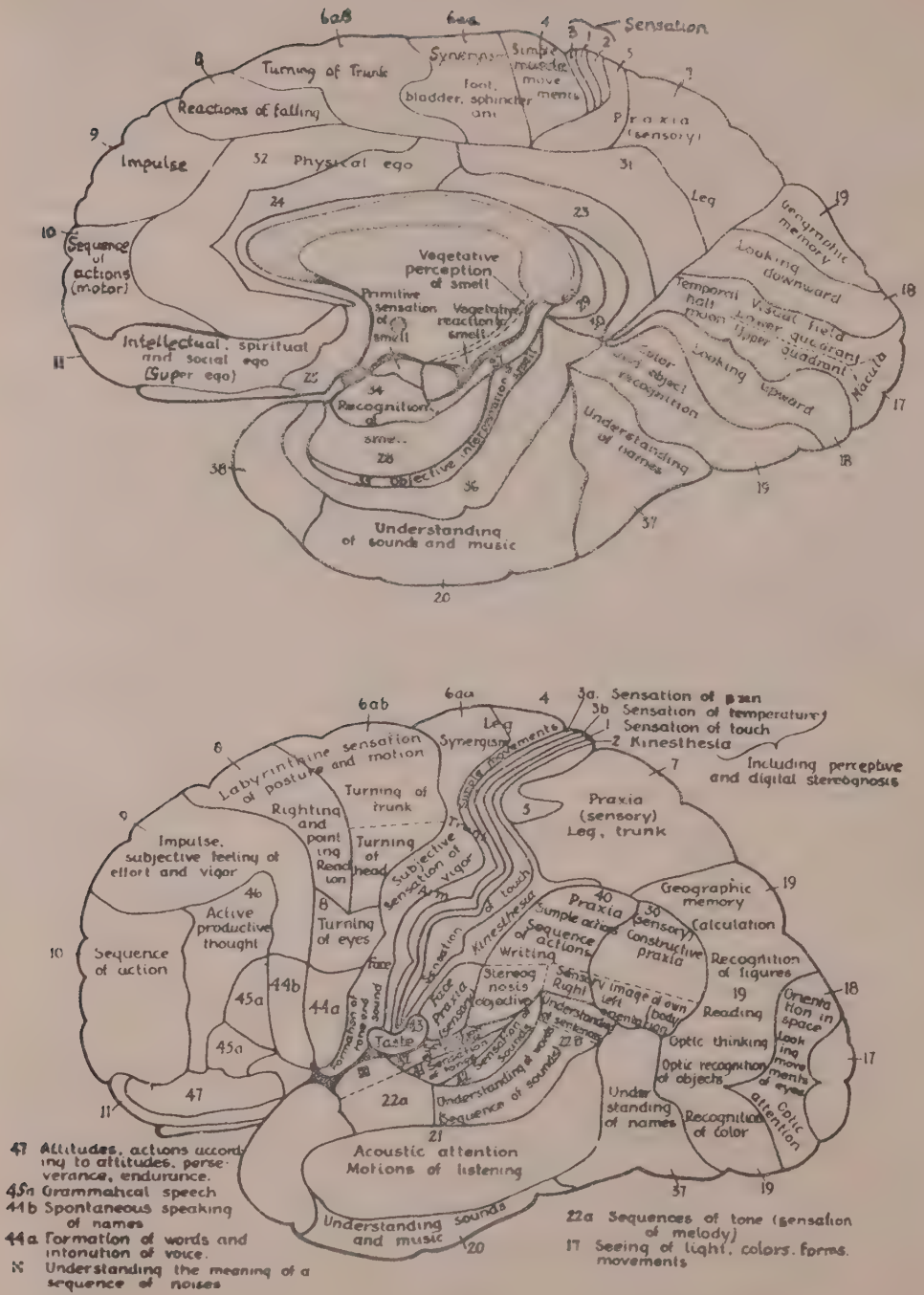
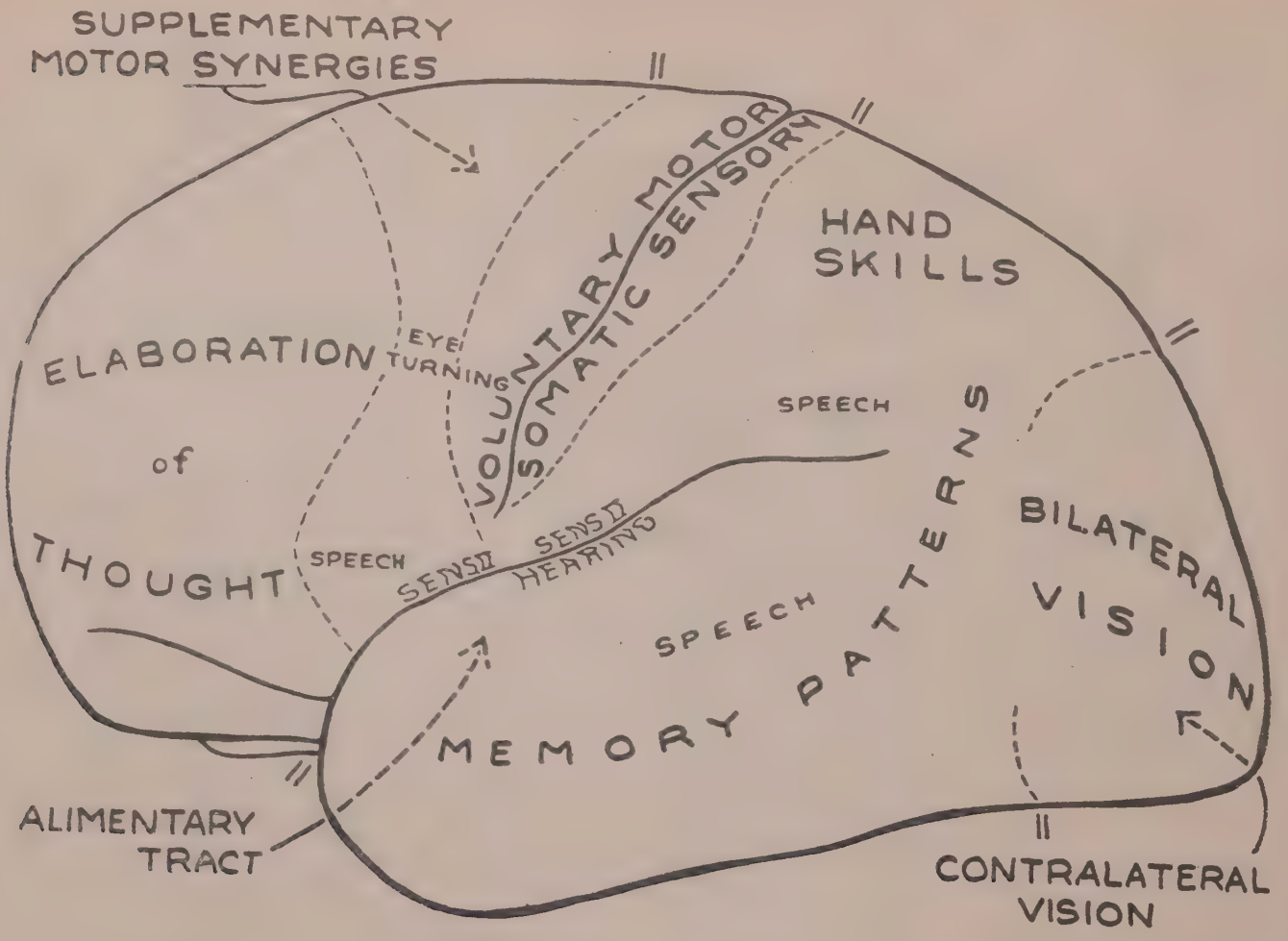


Fig. 1. (A AND B). Cerebral localization, in the main according to Kleist (9), adapted by Alexander (16).

Fig. 2. LOCALIZATION OF FUNCTION



By courtesy of Prof. Wilder Penfield

INFORMATION, MACHINES, AND BRAINS

by
A.M. Uttley.

(1) INTRODUCTION

This paper is concerned with the popular but dangerous task of comparing computers and brains. The only reason why such a paper should occur in a Symposium on Information Theory is that computers and brains appear to have certain common properties relevant to this subject.

(1) They require an input of information;

(2) They can emit information;

(3) They can store information;

(4) The input information may be partly ignored and partly transformed to form the output information.

However, the high speed computer did not come in the attempt to build a brain, although the aeroplane came from attempts to imitate birds; it came rather from continued development of a simple adding machine. Originally it was thought that such a machine could run through a series of instructions sequentially only, behaving rather like a gramophone. However, Babbage showed that instructions could be given in very general form and that the order of their performance could depend upon circumstances. As a further step the field of Mathematical Logic has proved amenable to treatment by machines. It now appears that machines can solve many problems hitherto considered to require "thought".

It is hoped, in this paper, to point out certain similarities of function, and of behaviour between computing machines and animals.

(2) FUNCTIONAL SIMILARITIES

Storage

An abvious similarity between brains and machines lies in their ability to store information; much has been written on the various ways in which this is possible. It will suffice here to point out that a very large part of behaviour is determined by simple storage of information.

Transformation of Information

This principle will be illustrated from a number of widely differing fields.

Mathematics

Consider a simple problem which can be solved by a computing machine.

$$(a_1 \pm \delta a_1)\, x + (b_1 \pm \delta b_1)\, y + c_1 \pm \delta c_1 = 0 \qquad (1)$$

$$(a_2 \pm \delta a_2)\, x + (b_2 \pm \delta b_2)\, y + c_2 \pm \delta c_2 = 0 \qquad (2)$$

where δa_1, δb_1, etc. are probable errors. There will be a solution:-

$$x = \alpha \pm \delta\alpha \qquad (3)$$

$$y = \beta \pm \delta\beta \qquad (4)$$

If the probable errors were zero, there would be infinite information; for finite probable errors it can be shown that the total information content of statements (1) and (2) equals that in (3) and (4). The process may also be considered as that of recoding.

As a second example consider the information laboriously collected by Tycho Brahé of the apparent motion of the planets. The same information was transformed into three short sentences or laws by Kepler; strictly speaking, they should have included statements of probable error. A law without such an attached condition contains infinite information about an infinite number of possible occurrences.

Logic

Consider the problem:-

Given that -

(1) All the communications I have received which were written on white paper were typed.

(2) No typewritten matter has been sent to me in an unsealed envelope.

(3) Some of my bills were on white paper.

(4) I have not paid any bills which did not come in unsealed envelopes.

Have I paid (a) all, (b) any of my bills?

Here again, of the initial information some is discarded and some is recoded to give the conclusion

(a) No;

(b) Unknown, insufficient information.

The objects discussed - bills - have five potential attributes whose presence or absence can be described by five binary digits. Set theory can be used to obtain the solution, which is determinate since this is a case of discrete statements without disturbances or probable error. The problem is soluble by means of a computer.

Civil Defence

Consider next the working of an A.R.P. system during raids. Certain information obtained at outposts is sent to a control centre. After various processes simple messages emerge of the form "Send Bill from A to B". Again no fresh information is created. The outgoing messages are framed according to quite definite coding rules which are applied to the input information. A new condition arises here in that part of the input information was known before the raid, for example, "A High Explosive Bomb Incident requires two stretcher parties". This a priori information learnt by past experience must be distinguished from a posteriori information of the coordinates and types of bombs fallen.

Living Organisms

Lastly one can point out in animal structure sense organs which receive new information, memory which somehow stores past a priori information, and a Central Nervous system which modifies information and emits instructions to motor organs.

(3) THE STRUCTURE OF A DIGITAL COMPUTER

No comments will be made upon any similarity of structure of brain and machine. A very short description is here included of the structure of a digital computer; this will help to make clear how certain forms of activity are possible for it.

A digital computer must contain the following elements:-

1. A store of digits. Binary digits are preferred, they are economical for arithmetical manipulation, and are suitable to represent "truth" and "falsity" in logical problems. Both numbers and instructions are stored.

2. A means of selection from the store. The selective organ is supplied with an "address" and makes available the contents of that "address".

3. In a comparison unit relations between two digits may be formed. If one digit is in the store a register is required into which can be placed the second digit, a second register is required into which the "relation" digit can be placed. Basically the "AND" relation and the operation "NOT" suffice to cover all logical and arithmetical comparisons.

The "AND" Comparison Unit operates as follows:-

Input A	Input B	Output
0	0	0
0	1	0
1	0	0
1	1	0

This may be written more neatly

B \ A	0	1
0	0	0
1	0	1

The operation "NOT" involves only one input so is not a comparison; it has the property:-

Input	Output
0	1
1	0

The "Identity" Comparison Unit has the property

B \ A	0	1
0	1	0
1	0	1

4. An instruction register is required into which is placed from the store an instruction or set of digits which define:-

(a) The address of the number now required from the store.

(b) The operation now to be performed. By a suitable code the operation is described as to

(i) Route of digit

(ii) Direction of transfer - to or from store

(iii) Type of relation required.

5. A control register contains the address of the next instruction. Normally this register contains a steadily increasing positive integer, being increased by 1 at the end of each operation; in this case instructions are performed sequentially.

6. Conditional Control. It was Babbage who considered the effect of overruling the existing contents of the above register by placing in it a new number from the store IF a certain condition arose. (A convenient condition is that the contents of a chosen register must be negative.) It is this step which has so extended the scope of automatic computers; it consists of overall feedback by which the results of actions in the computing organ cause changes in the behaviour of the whole machine.

7. Input and Output Organs. Lastly a computer requires a supply of information or new material upon which to work. Instructions as to what shall be done with this information may be partly stored already in the machine and partly supplied as fresh information.

The output organ is a means of display of the manipulated information.

(4) SIMILARITIES OF BEHAVIOUR

Similarities will be considered between animal behaviour and that of a series of machines of different type.

Gramophone Activities

A gramophone with a means of selecting records by means of a pre-arranged code can behave in a way that resembles much human activity, such as the playing of a musical score without modification by relevant stored past experience, if any; and the reproduction of learned bookwork in examinations.

Servomechanisms

Here behaviour is determined by a priori information stored in the form of a built-in structure, the animal analogue is the instinctive behaviour pattern. The a posteriori information enters through a sensitive detector which converts energy from its original form to another. At any point in this input channel if energy is introduced which is of the same nature as that for which the channel is designed, then the mechanism will react as if the original form had fallen on the detector. Analogous phenomena in the animal world are "equivalent stimuli", and "illusions".

In all such activities, from that of a Selsyn Repeater when a shaft rotation is repeated in a remote shaft, to an automatically controlled oil refinery; from a knee jerk to the courtship of birds; there is a correspondence between stimulus and response. Nevertheless in both fields disturbances produce variations of the basic behaviour.

Computers

A computer has the ability to store, compare, and to decide. It can learn by rote and by experiment. A growing and accessible library of programmes constitutes a growing body of past experience. In the field of mathematics the library will be given routines or abilities of growing complexity such as:-

(1) Multiplication, division

(2) Tables of functions

(3) Solution of equations.

Past Experience. As in the examples of Civil Defence and of the animal, a computer can contain a priori information. A computer with nothing in its store is more helpless than a new-born babe, but a computer has the most important property of faultlessly learning by rote and experience; as to the first method, once instructions are stored to deal

with a certain situation, then the machine will carry out the correct procedure for all such future situations; as to the second method, it is the conditional instruction which is vital since it allows judgement and alternative actions. The first time such an alternative exists, a single trial with the resulting success or failure will make it possible for a computer to change the associated conditional instruction into a certainly correct unconditional instruction. The computer has learnt from a single mistake.

As an example of learning by experience one can consider the learning of a maze. As posed to a digital computer the maze can be described by an array of binary digits where "1" can represent "wall", and "0" represent "path". A routine of instructions can be introduced which will apply generally to any maze. The conditional instruction (it has often been called a branching instruction) will be required at each junction of the maze. There will be two forms of failure:-

(1) I reach a cul-de-sac.

(2) I have been here before.

The second failure can only be recognised if the computer records all its previous moves; so too a man requires memory in order to detect the second type of failure.

For either type of failure the programme of instructions causes the branching instruction to be converted to an unconditional one, the right one, so that the next time the maze is entered the path to the centre will take less time. With a perfect memory the computer will make no mistakes this second time.

If the maze contains branches neither of whose paths result in failure but one of which is shorter, then an addition to the programme will enable the computer to discover a shorter path. An imperfect memory with only certain probabilities of recording a failure will cause the time of threading the maze to fall exponentially to a minimum.

Pattern Recognition. One may now consider pattern recognition which is a complex example of the identity relation.

The ability to detect relations has been held to be a test of intelligence and yet the computer contains a Comparison Unit for this very purpose.

Consider a simplified form of common intelligence test.

Four objects on each described by five binary digits, accordingly as they are or are not in five independent classes. In the four arrays of binary digits is there any common pattern? Are some arrays not of this pattern?

A_1 is 1 1 1 0 0

A_2 is 1 1 1 0 1

A_3 is 1 1 1 1 0

A_4 is 1 1 0 1 1

The IDENTITY comparison unit is capable of taking two of the arrays or numbers and comparing them digit by digit emitting a 1 for identity and a 0 for difference. Comparing A_1 and A_2 we obtain the relation number 1 1 1 1 0.

Comparing A_1 and A_3 we obtain the relation number 1 1 1 0 1.

Comparing these two new numbers by means of the "AND" relation we obtain a third number 1 1 1 0 0. So that the three numbers A_1, A_2, and A_3 are of the class 1 1 1 0 0. A_4 is not of this class. Similarly, A_2, A_3, A_4 are of the larger class 1 1 0 0 0 which contains class 1 1 1 0 0. A_1 is in this larger class. Also, A_3, A_4, A_1, are of this class 1 1 0 0 0 as are A_4, A_1, A_2. And so the unambiguous result is obtained. The number 1 1 1 0 0 defines what aspects shall be attended to. The combination

1 1 1 0 0
1 0 1

might be called a Universal.

The above process is one of inductive reasoning, of perceiving a general pattern or law from a series of particular observations. It is also one of information transformation. One is led to the conclusion that inductive reasoning does not produce new information; this is in disagreement with the quotation in E.C. Cherry's paper from Dr. R.A. Fisher's book "The design of Experiments", namely "Inductive inference is the only process by which new knowledge comes into the world".

It is suggested that:-

1. Inductive reasoning is possible for a computer.

2. Inductive reasoning produces no new information if the resulting pattern or law is limited to describing only the actual observations made by experiment.

3. Extrapolation or assumption that a law applies to circumstances not yet observed is an act of scientific faith. It produces no new knowledge unless confirmed by experiment.

All these abilities have come to the computer from the external world, which includes the designer of the machine and its programmes. Except in one respect, this statement applies equally to the animal and to man. The exception concerns the creation of new information in and by the organism. Creative activities will be discussed later. It is important that the exception is not concerned with predictability. It is easy to make a machine unpredictable without any creative facility. This can be done in two ways. Firstly one can flood the machine with input information from a complex external environment. This is not true unpredictability, it is just that the calculations would take a long time. The "tortoise" of Dr. Grey Walter exhibits this class of behaviour, IF friction and electronic noise are at a negligible level. This condition leads to the second method of producing unpredictability, that of introducing a random process somewhere in the system. This step in itself is not very clever, but the combination of random behaviour with the existing ability to compare the results of one activity with another is a powerful tool with which to improve learning by experience; it may be seen to operate when an animal is placed in a strange environment; for example when a cat is placed in a cage and must discover the mechanism of a latch on a door.

In terms of a computer we are speaking of the ability of a machine to construct its own programmes on a trial and error basis. The influence of the external environment becomes less obvious; it has been at the high level of inserting a programme to cause the machine to make a random change in a programme and to reject the less satisfactory of the two alternatives.

Animal Behaviour

It is suggested that two steps may be taken towards the quantitative analysis of animal behaviour.

Firstly, the relevant environment can be analysed into a numberable set of elementary propositions. For example, there is no mystery as to why animals and men find it harder to recognise Fig. 1a from Fig.1b (Ref. 1). There is more metrical information in Fig. 1a. To put it

another way, in order to communicate Fig. 1a one must supply the four pairs of coordinates of the four points, say $(x_1\ y_1)$; $(x_2\ y_2)$; $(x_3\ y_3)$; $(x_4\ y_4)$. To communicate Fig. 1b it is only necessary to communicate the quantities $(x_1 \pm x_2;\ y_1 \pm x_2)$. The information ratio is 8/3.

Secondly, the manipulation of information can be broken down into its logical elements of binary digit storage, and binary digit comparison. In the example chosen above of pattern recognition, there were 100 binary storages, and 80 binary comparisons.

One could discover whether tasks measured this wasy correlated with brain volume, for example.

Creative Activities. There seem to be three possible views of the meaning of creation.

(1) Causeless occurrence. Example: Spontaneous Creation.

(2) The unperceived cause. Example: Discovery of Scientific Law.

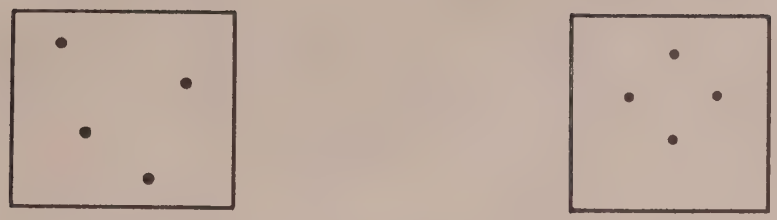
(3) The unknown case. Example: The impulse to do good.

For "cause" we may write "information flow".

It is believed that the first idea is not now accepted in any field of thought. In the second view any inexplicable human output is to be called a creation; so that Keplers' work would be the result of creative thought. Yet all the information was already provided. It remained to perceive the relations. In the view here taken a relation between two given entities is not regarded as new information but as a transformation of the original input information. It is in fact an essential part of the scientific view that anyone could have discovered Kepler's Laws, they are independent of the scientist. Nevertheless the pyramid of relations between relations could be detected more readily by one man than another, and this is possible for a machine. In the same way, a product of the imagination arises from remembered percepts and perceived relations between them. One man will walk the streets and see no reason between himself and a cloud overhead. Another man will see the relation and write the words "I wandered lonely as a cloud". This too seems not impossible for a machine.

Thirdly there may be an unknown cause operating in man, which we have not and perhaps cannot introduce into a machine. Julian Huxley in his book "Religion Without Revelation" would have none of this. Many will agree with him. But if of two men one has experienced such information flow and another has not, no discussion will unify their views.

FIG. 1.



PATTERN RECOGNITION.

STATISTICS FOR THE CHESS COMPUTER AND THE FACTOR OF MOBILITY

by
Eliot Slater

Shannon has argued that the problem of providing a programme for a chess-playing computer is of theoretical interest, and its use might lead to a wide range of practical developments. The problem is also interesting psychologically. If the human and the mechanical players are to play the same game, they will each have to be directed by concepts which have a certain equivalence. But the concepts used by the skilled human chess-player are both subtle and complex, and for the purpose of programming a computer they will have to be reduced to their simplest form. Chess-masters are, as a class, men of considerable general intellectual ability, and come from the ranks of professional men, mathematicians, scientists, lawyers, etc. They have in addition a special ability. Very few chess-masters, who began the game early, did not show unusual excellence at it at a very early age. The specific chess ability begins to show itself, given the opportunity, at about the age of eleven. Furthermore, there are few, if any, chess-masters who cannot play blindfold, and play many games at once, achievements which are entirely beyond the powers of the ordinary player. The order of intellectual activity which we are required to reduce to simple terms is therefore of a superior kind.

Shannon has suggested that one practical strategy for the chess computer would be for it to combine a deeper analysis of forcing variations with a two- or three-move analysis of development in more quiescent positions. Both are necessary if the machine is neither to fall into simple tactical traps, nor to fail to develop its pieces in an adequate way. The present investigation is confined to the latter aspect of play. Statistical data have been taken from the games played in the following tournaments: St.Petersburg 1914, New York 1916, London 1922, Hastings 1922, Kecskemet 1927, Capablanca-Alekhin match 1927, Alekhin-Euwe match 1933.

It would simplify the problem of programming the computer if it could be provided with only a few parameters for the measurement of a given chess position. The total value of a position would then be the weighted sum of the values of these parameters, i.e. the directive for the machine would be a discriminant function which could be calculated by the statistical analysis of chess positions of known value. Shannon has given a list of fifteen features which should be evaluated by the machine. It is submitted that this list could be simplified. The three noteworthy parameters appear to be those of material, structure and function. Structure is inherent in the pawn formation, which changes only slowly during the course of a game. It is far from easy to provide any single scale of measurement for its evaluation. Material, on the other hand, is easily measurable. For practical purposes we can adopt Shannon's scale of 9 for Q, 5 for R, 3 for B or S, 1 for P. If the value of a game is +1 in the case of a win for White, 0 in the case of a draw, -1 in the case of a win for Black, then the value of a chess position of a defined kind, Shannon's f(P), is $(w - b)/(w + d + b)$, where w, d and b are respectively the number of White wins, draws and Black wins which have been obtained from that kind of position. From the games analysed an advantage of one pawn has at the 20th move if positive i.e. in favour of White, a value of +0.27, if negative, i.e. in favour of Black, a value of -0.21.

Material gains are therfore, as we could have guessed, highly decisive. They are, however, attained as a rule only after a considerable number of moves and with difficulty. Out of 350 master games, only in 98 cases had a material advantage been won by the 20th move. This factor by itself will be of value to the machine towards the end of the game in administering the coup-de-grace, and in ordinary situations in preventing it from falling into crass error. For move-to-move guidance, advantages measurable in smaller units will be evaluated. This, the strategic advantage, is involved in the functional factor.

This factor appears to be largely identical with mobility. Its value is recognised by chess experts, but it is often inaccurately judged. On the simplest view it is fundamental. The outstanding characteristic of a chess position is that it is a dynamic situation with a definite and limited number of degrees of freedom (the total number of legal moves). The game cannot be finally won without administering mate, when the opponent's degrees of freedom are reduced to zero. A series of checks, known by chess-players to be highly dangerous, is so because it reduces the opponent's mobility to the lowest possible level short of paralysis, and often eventually compels a disastrous move.

If the degrees of freedom available to each opponent are charted through the course of a game, a double curve is obtained whose rises and falls correspond closely with the strategic advantage. Means taken from 78 arbitrarily selected games which ended with a decision on or before the 40th move shows the following state of the parties:

After move	Winners	Losers
0	20.0	20.0
5	34.2	33.9
10	37.5	36.0
15	39.7	35.2
20	38.9	36.4
25	39.6	31.9
30	35.6	27.7
35	31.7	23.2

These figures show (1) a rise in the degrees of freedom as pieces are developed, followed by (2) a fall as pieces are exchanged, and (3) a consistent and increasing difference to the advantage of the winners.

Shannon's f(P) has as its limits +1 and -1. We can define an advantage in mobility, M, as $(dF_W - dF_B) / (dF_W + dF_B)$, which has the same limits. It has been found from a total of 380 games from the tournaments mentioned above, i.e. including all those which proceeded as far as the 20th move without one or other player having won an advantage in material, that there is a close relation between f(P) and M. This is shown in the following table:

Value of M at 20th move	f(P)
+ .26	+ .46
+ .19	+ .35
+ .14	+ .27
+ .09	+ .17
+ .04	+ .09
- .01	- .05
- .07	- .06
- .13	- .03
- .24	- .43

If graphed these points approximate to a straight line, and to the equation $f(P) = +.086 + 1.658M$.

We may use the value of M to predict the value of f(P), and are then interested to know its efficiency. If wins for White, wins for Black and draws are in equal proportions (as is very nearly the case), a random classification of a series of games will result in a misclassification rate of 89 points per hundred. If games are ranked by the value of M at the 20th move, the misclassification rate drops to 63 points per hundred, and if both the values of a material advantage and of M are combined by simple addition, the rate drops still further to 55 points per hundred. Efficiency is therefore rather low. It is rather doubtful whether it could be much improved without the intrusion of a subjective element. The players themselves "misclassify" their

games, as it is far from infrequent for a player to lose from a theoretically drawn position, even in a fairly advanced stage of the game, or to draw or lose from a theoretically won position. The frequency with which this is actually observed depends on the care and skill of the subsequent analysis to which tournament games are usually subjected when play is over. Much depends, therefore, on the annotator. With a highly skilled and interested analyst like Alekhin, one finds that, in his judgement, the "misclassification" produced by errors in play approaches in the tournament games he has annotated to 40 points per 100.

To summarise the effect of these arguments, it does seem possible that a chess computer which was programmed, beyond immediate tactical tasks, to maximise the mobility difference between itself and its opponent over a series of moves, might play a strategically tolerable games of chess. The mobility factor by itself might be a sufficient measure of all the factors (Rook on open file, etc.) listed by Shannon as requiring evaluation under his headings (3) and (5). Moreover the concept of mobility overlaps with other concepts much used by chess players. Control of space, when defined as the number of squares in the opponent's half of the board whose use can be denied him for at least one move, proves to be correlated with mobility by a coefficient of +0.83, i.e. the two concepts are nearly identical. "Development" is largely tantamount to the acquisition of increased mobility for all pieces together and for each piece separately. "Initiative" is nearly always in the hands of the player with the advantage in mobility, and "having the attack" seems to consist in having the opportunity for its aggressive application. "Combination", though nowhere very precisely defined by chess authors, seems to be the same thing as the exchange of one sort of advantage for another, e.g. when superior mobility enables a rapid concentration of attack on a piece which cannot be so quickly defended, or when a sacrifice in material buys an open file bearing on the opposing King.

The speculation may be offered that many other games, from draughts to war, may be found by appropriate analysis to involve the same concept, i.e. that advantage lies in creating a difference in mobility. For the more restricted problem of programming a chess computer, however, these preliminary investigations also suggest that a digital computer may prove an inefficient instrument, and that an analogue machine, or a combination of digital and analogue machines, will provide better results.

REFERENCE

Shannon, C.E.: "Programming a Computer for Playing Chess". Phil.Mag., 41, 256-275 (1950).

CRITERIA OF PREDICTION AND DISCRIMINATION

by
J. H. Westcott

INTRODUCTION

The topic of predication has an air of witchcraft about it that should be renounced at the outset. This paper is not concerned with the miraculous, but the realisation of the possible. In the era of determinism, absolute prediction of events seemed a possibility and only ignorance of detail a bar to its universal realisation. The field of astronomy provided encouraging support of the hypothesis. To-day it is realised that some physical phenomena cannot be adequately accounted for on a basis of Newtonian mechanics and for these interesting cases where interaction effects are important ignorance of detail is in the nature of things and permanent. It is thus necessary to think in terms of statistical mechanics (after Gibbs).

It is now established and widely recognised that a valid theory of information must be based upon a statistical analysis of signals or events, and in the presentation of this paper this fact will be found in the background to make itself evident at several critical points. This study is concerned with the history of waveforms, best described as fluctuations, upon which linear operations are performed. It will be useful to think of the fluctuations as composed of sets of transients commencing at randomly-spaced points in time and having random arguments (or sizes). An aspect of information in a time series may be usefully emphasised here. The component transients are continuous waveforms and so have the apparent potentiality of conveying infinite information; but this aspect is not at issue. The particular continuous transient may be regarded as conveying information in a structural fashion; its very form being its significance.

With this explanation the problem of predicting such a fluctuation may be posed thus: given a sufficiently representative sample of signals known to be in statistical equilibrium, to what extent is it possible to devise a weighting function which when acting on the waveform gives the best forecast of the future of the fluctuation; the weighting function to be such that it can be physically realised by a stable network of linear elements? The critical part of the process as far as the present discussion is concerned is the matter of how to define and specify "best". A natural extension of the prediction problem, although somewhat more difficult than it, is that of discrimination; that is the "best" recovery by the linear operations of a stable network of a message corrupted by the presence of disturbances.

Much of the material of this paper is speculative and greater liberties have been taken than is normally allowable under such distinguishable auspices with the hope of provoking discussion from this talented assembly.

DETERMINISTIC PREDICTION

Taylor's Series was common knowledge among mathematicians by the year 1720. The history of prediction as a logical possibility dates from this period. As an example of determinist prediction consider the application of this series to a single transient $f_1(t)$ which is zero for $t < 0$ but otherwise continuous and differentiable. By Taylor's Series we have:-

$$f_1(t+\alpha) = f_1(t) + \alpha f_1'(t) + \frac{\alpha^2}{2!} f_1''(t) + \text{------}$$

where α is the prediction time.

The series wherever terminated has a remainder term; in favourable cases of smooth curves the remainder will be small or even zero after a few terms, but error is introduced by the neglect of higher terms unless these are fortuitously all zero. The accurate realisation of even the first few differential coefficients of a curve presents difficulties in practice and error arises here. Attempts at prediction have been made on this basis with some success in the case of the simpler smooth curves, but an assessment of the sources of error in the determination or omission of coefficients is difficult to achieve. It was, however, confidently supposed that if accuracy in this matter of the differential coefficient could be achieved absolute predication would be accomplished. It is interesting to make a preliminary excursion to discover to what extent this was physically realisable:-

$$f_1(t+\alpha) = f_1(t) + \alpha f_1'(t) + \frac{\alpha^2}{2!} f_1''(t) + \cdots$$
$$= (1 + \alpha.\lambda + \frac{\alpha^2}{2!} \lambda^2 + \cdots).\ f_1(t)$$
$$= e^{\alpha\lambda}.\ f_1(t) \qquad (1)$$

(where $\lambda \equiv \frac{d}{dt}$ is an operator that commutes with constants satisfying the laws of algebra so may be treated as an algebraic quantity).

This is the symbolic form of Taylor's series. Thus $e^{\alpha\lambda}$ (where α is +ve) may be regarded as the ideal prediction operator; it is independent of the waveform it operates on, always giving a shift forward in time of α seconds. It summarises in a single operator all the differential coefficients with their respective arguments of the Taylor series. It is ideal but unfortunately not physically realisable as a stable network.

PHYSICALLY REALISABLE OPERATORS

It is now necessary to specify what class of operators can be realised as stable physical networks. A network is stable if for every bounded input it produces a bounded output; when this is not the case the network is unstable and unusable. If $h(t)$ is the response of the network to a unit mathematical impulse this condition for stability is satisfied when

$$\int_0^\infty |h(t)| \, dt$$

is bounded. But the Impulse response is the sum of the natural modes of the network:-

$$h(t) = \sum_{\kappa=1}^{n} a_\kappa e^{+\lambda_\kappa t}$$

where $e^{\lambda_\kappa t}$ are the natural modes of the network (λ complex). Clearly $h(t)$ will satisfy the condition given provided no λ_κ has a positive real part. But the λ_κ's are the poles of the transfer function of the network $H(\lambda)$ where $H(\lambda)$ is the Laplace Transform of $h(t)$:-

$$H(\lambda) = \int_0^\infty h(t)\, e^{-\lambda t}\, dt. \triangleq L\{h(t)\}$$

Thus the network is stable provided the operator $H(\lambda)$ has no poles on the imaginary λ-axis or in the right half λ-plane. It should be added that it is possible to give a rigorous proof that the network is stable when $H(\lambda)$ is analytic in the right half-plane and well-behaved on the imaginary axis (1). Since the operator about to be considered has an essential singularity in the right half-plane the existence of such a proof is strictly speaking essential to the argument. This condition then of being analytic in the right half-plane and on the imaginary axis is the one that must be imposed upon operators if they are to be realisable as stable physical networks.

Returning to the ideal prediction operator $e^{\alpha\lambda}$ (α positive) it is clear in the light of the condition just stated that this operator does not belong to the class of stable physically realisable networks. So absolute prediction cannot be achieved by the use of stable networks. But why restrain the condition to stable networks? - unstable networks are very easy to realise as anyone with experience of closed-loop systems knows to his cost and in a purely determinist scheme of things an unstable network, once conceived, would not necessarily go into immediate paralytic instability and might thus be persuaded to give pure prediction of a continuous uncorrupted waveform. Unfortunately an unlooked for random disturbance of thermal agitation is always in practice present and this uninvited representative of the second law of thermo-dynamics makes the restriction of operators to stable networks the only practical course. Ignorance of detail in thermal agitation is inevitable and that seeming possibility of determinist philosophy, of absolute prediction, a myth; at least from the viewpoint of reality that our particular nitch in the universe presents.

The question arises as to what can be salvaged from the ravages of indeterminacy. Consider the situation represented by Fig.1 in terms of operators:- a waveform $F_1(\lambda)$ passes through a network whose transfer function is $H(\lambda)$ and is required to have an output $e^{\alpha\lambda}.F_1(\lambda)$

$$\left[\text{where } F_1(\lambda) = \int_0^\infty f_1(t)\, e^{-\lambda t}\, dt. \triangleq L\left\{f_1(t)\right\}\right]$$

now $\int_0^\infty |f(t)|dt$ is bounded so that $F_1(\lambda)$ is bounded in the r.h.p. (right half λ-plane) and $\int_0^\infty |h(t)|\, dt$ is to be likewise for $H(\lambda)$ a stable network, so that the product $F_1(\lambda).\, H(\lambda)$ (the result of $F_1(\lambda)$ passing through the network) will also be bounded r.h.p. Thus a valid solution to the problem occurs provided that part of $\left[e^{\alpha\lambda}.F_1(\lambda)\right]$ is taken that is so bounded:-

$$F_1(\lambda).\, H(\lambda) = \left[e^{\alpha\lambda}.F_1(\lambda)\right] \quad \text{bounded r.h.p.} \tag{2}$$

$$\text{but } e^{\alpha\lambda}.F_1(\lambda) = \int_0^\infty e^{-\lambda t}.f_1(t+\alpha)\, dt + \int_{-\infty}^0 e^{-\lambda t}\, f_1(t+\alpha)\, dt \tag{3}$$

where the first integral on r.h.s. of (3) is bounded r.h.p. only and the second integral is bounded l.h.p. only.

thus (2) becomes $F_1(\lambda).H(\lambda) = \int_0^\infty e^{-\lambda t}.f_1(t+\alpha)\, dt$

$$\text{or } H(\lambda) = \frac{1}{F_1(\lambda)} \int_0^\infty e^{-\lambda t}.f_1(t+\alpha)\, dt \tag{4}$$

which gives the required stable realisable operator.

It will be noted that the operator is now dependent upon the waveform it operates on.

The operation that the network performs is clear from Fig.1. Before the start of the transient the network can do nothing (otherwise it has the gift of prophecy). At the commencement of the transient the network causes the output to be a replica of the transient as it will be α seconds later from that point onwards, in time, following the transient perfectly and exactly α seconds ahead. The shaded area must be lost and is represented by the second integral in (3). The answer finally obtained is a relatively simple one and not entirely unexpected, but nevertheless in the getting of it the very foundations of determinism have been plumbed and found wanting.

STATISTICAL PREDICTION

So far the determinist approach to prediction has been exploited as fully as physically possible. The determinist view-point practically rules out the possibility of receiving information. This emerges clearly from the previous papers, where it has been shown that the element of a priori choice is essential for the effective transmission of information. It is now necessary to take the statistical nature of signals into account; to establish what degree of orderliness may be reasonably assumed in the long run, if not in the particular case; to find how the determinist view may be modified to be in keeping with the more mature view of the nature of the signals.

In the first instance consider a set of transients of the same form as $f_1(t)$, but of randomly varying arguments (sizes) and randomly spaced in time. Thus the structural, or à priori, nature of the signal is specified and singular. Unknown, are the magnitudes (metrical nature) and the time origins. In this case the physically restricted determinist solution is suitable for direct application to each transient individually and so to the sequence of transients collectively giving rise to an error which will be proportional to the shaded area of Fig.1 for the complete waveform. A simple example of such a case has been carried out experimentally in the U.S.A. for a signal consisting of the output from a damped tuned circuit that has random noise impressed on it. Predictions obtained and recorded photographically show a satisfying quantitative agreement with the expected degree of error. (2)

No question of a criterion of prediction arises yet, since all that can be physically realised has been, and that only in an ideal meanner. The situation is modified when one considers approximating to the required operator by a network of physical elements known to have certain limitation (such as the inevitable resistance of inductances coils).

The next stage is to assume that the transient waveforms are not all alike. The first and most obvious method of dealing with this case (for there is often merit in simplicity) is to take the average in time of all the transients or a representative sequence of them and to use this average representative transient as the basis for the assessment of the best physical operator after the previous determinist manner. Note that the representative transient is not usually a member of the set of transients. Use has been made here, unobtrusively perhaps, of an ergodic hypothesis; which is a way of saying a necessary assumption of statistical reasoning has been presumed about the set of transients. Instead of considering a particular set of transients one considers a class of sets of transients such that for each member of the class the average transient is the same. In certain conditions this ergodic relation can be proved as a theorem (3). It may be shown that the mean square deviation between the transients and a set of representative transients obtained on this averaging basis in time is a minimum if the waveforms have zero mean.[x]

There arises the practical difficulty of how the averaging process may be carried out and the representative transient derived given, as happens in practice, a set of transients $f_n(t)$ as an assembled waveform $f(t)$.

---

x To statisticians apologies are due for this gross simplification of their art.

Consider first a single transient $f_1(t)$:-

$$\lim_{T\to\infty} \int_{-T}^{T} f_1(t)\, f_1(t+\tau)\, dt. = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_1(j\omega).\overline{F_1(j\omega)}\, e^{j\omega\tau} d\omega. \quad (\omega \text{ real})$$

where $F_1(j\omega)$ is the Fourier transform of $f_1(t)$: by the complex multiplication theorem. Thus for $f(t)$ the assembled waveform

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f(t).f(t+\tau)\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\omega).\overline{F(j\omega)}.e^{j\omega\tau}\, d\omega = \phi(\tau).$$

where $F(j\omega).\overline{F(j\omega)} = \lim_{T\to\infty} \frac{1}{2T} \sum^{n} F_n(j\omega)\, \overline{F}_n\, (j\omega)$ is known as the spectrum intensity function of the waveform and is its average spectral density distribution with respect to frequency. $\phi(\tau)$ is the log-covariance of the waveform and is a practically convenient quantity to measure. Its Fourier transform is the spectrum intensity function.[xx]

The representative transient squared $f_r^2(t)$ is that transient whose Fourier spectrum is the same as the spectrum intensity function of the complete waveform $f(t)$. Since the phase relations between component vibrations have been lost in this process there will be an infinite number of such transients. There is, however, a unique one having a minimum-phase property which is chosen for simplicity. It is now a small step to determine the best physically realisable and stable operator by performing the previous operation (eqns. 2 - 4) with respect to the representative transient thus

$$H(j\omega) = \frac{1}{F(j\omega)} \cdot \int_{o-}^{\infty} e^{-j\omega t}\, f_r(t+\alpha)\, dt \qquad (5)$$

(where ω may be assumed complex) which always gives a stable realisable operator when $F(j\omega)$ is chosen as the minimum phase spectrum. In terms of $\lambda = j\omega$ by change of variable:-

$$H(\lambda) = \frac{1}{F(\lambda)} \int_{o-}^{\infty} e^{-\lambda t}\, f_r(t+\alpha)\, dt \qquad (6)$$

where α (a positive quantity) is the prediction time.

A word about the errors accumulating with the use of such a predictor is needed. By eqn. (3) it is seen that

$$e^{\lambda\alpha}.F(\lambda) = \int_{o}^{\infty} e^{-\lambda t}\, f_r(t+\alpha)\, dt + \int_{-\infty}^{o} e^{-\lambda t}\, f_r(t+\alpha)\, dt \text{ on the average.}$$

---

xx When $f(t)$ is supplied by a voltage source across a 1 ohm resistance the power dissipated may be expressed in two ways:-

$$\underbrace{\frac{1}{2\pi}\int_{-\infty}^{\infty} |F(j\omega)|^2\, d\omega}_{\text{integrating the spectrum intensity function w.r.t. frequency}} = \lim_{T\to\infty} \frac{1}{2T}\underbrace{\int_{-T}^{T} |f(t)|^2\, dt}_{\text{integrating the mean energy w.r.t. time}} = \phi(0) \text{ watts}$$

Thus the error arising when $e^{\lambda\alpha}.F(\lambda)$ is replaced by eqn. (6) is connected with:-

$$\int_{-\infty}^{0} f_r(t+\alpha)\,dt = \int_{-\infty}^{\alpha} f_r(t)\,dt = \int_{0}^{\alpha} f_r(t)\,dt \text{ since } f_r(t) = 0;\ t < 0$$

on the average. That is to say the mean square deviation in prediction is given by:-

$$\int_{0}^{\alpha} |f_r(t)|^2\,dt \qquad (7)$$

This is analogous to the square of the shaded area in Fig.1.*

DISCRIMINATION

The case of the predictor has served to show that a simple and elementary criterion for assessing the operations of a network on a random waveform is to minimise the mean square deviation of error. On this basis it is a simple matter to carry the process a stage further to deal with the more immediately useful problem of realising a stable network to discriminate between a wanted message and an unwanted disturbance. A fundamental property of information as transmitted through a communication channel is that it becomes corrupted and so lessened in significance but can never become of increased significance. When such a corrupted signal has been received an important problem is that of recovering as much as possible of the original message. The desideratum is a stable network that minimises the mean square deviation between a corrupted signal and the original message. The diagram (Fig.2) presents the problem.

$g(t)$ is the message whose spectrum intensity is $G(j\omega).\overline{G(j\omega)} = |G|^2$.
$n(t)$ is the disturbance whose spectrum intensity is $N(j\omega).\overline{N(j\omega)} = |N|^2$.
$g(t)$ and $n(t)$ are assumed incoherent; $H(j\omega)$ is the network transfer function.

Let α be the time shift in passing through the network (a delay when α is negative as in fig.2).

The variable ω may be considered as a complex variable and when ω is complex $\overline{F(j\omega)}$ means $F(-j\omega)$

The mean square deviation between message and output is given in time language by:-

$$I = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \left\{ \left| g(t+\alpha) - \int_{0}^{\infty} h(\sigma).g(t-\sigma)\,d\sigma \right|^2 + \left| \int_{0}^{\infty} h(\sigma)n(t-\sigma)d\sigma \right|^2 \right\} dt$$

since $g(t)$ and $n(t)$ are incoherent; or in frequency language:-

$$2\pi I = \int_{-\infty}^{\infty} \left\{ \left| e^{j\omega\alpha} - H(j\omega) \right|^2 |G(j\omega)|^2 + |H(j\omega).N(j\omega)|^2 \right\} d\omega.$$

$$= \int_{-\infty}^{\infty} \left\{ \left(1 - e^{j\alpha\omega}\,\overline{H} - e^{-j\alpha\omega}.H\right) |G|^2 + |HF|^2 \right\} d\omega.$$

since $|G|^2 + |N|^2 = |F|^2$

---

* It may be confessed that the use of a set of transients in this analysis is merely to simplify the task of thinking about the problem but is not essential to the development. Although many practical problems arise in this form, many that do not are equally amenable to the application of the method.

Let $L(j\omega) = \frac{|G(j\omega)|^2}{F(j\omega)}$ then adding and subtracting $|L|^2$:-

$$2\pi I = \int_{-\infty}^{\infty} \left\{ |G|^2 - |L|^2 + |L|^2 - \left(e^{j\alpha\omega}.\overline{H} + e^{-j\alpha\omega}.H \right) |G|^2 + |HF|^2 \right\} d\omega$$

$$= \int_{-\infty}^{\infty} \left\{ \frac{(|F|^2 - |G|^2)\,|G|^2}{|F|^2} + |L.e^{j\alpha\omega} - H.F|^2 \right\} d\omega .$$

$$= \int_{-\infty}^{\infty} \left| \frac{N.G.}{F} \right|^2 d\omega + \int_{-\infty}^{\infty} |L.e^{j\alpha\omega} - H.F|^2 \quad d\omega$$

The first integral is independent of $H(j\omega)$ and so is an irreducible error.

The second term is minimised for $H(j\omega)$ a stable network when:-

$$H(j\omega) = \frac{1}{F(j\omega)} \left[L(j\omega).e^{j\alpha\omega} \right]$$

bounded lower half ω -plane or by change of variable $\lambda = j\omega$:-

$$H(\lambda) = \frac{1}{F(\lambda)} \cdot \left[L(\lambda).e^{\alpha\lambda} \right]$$

bounded r.h. λ -plane

$$= \frac{1}{F(\lambda)} \int_{o}^{\infty} e^{-\lambda t}.1(t+\alpha).dt \qquad (8)$$

where $1(t+\alpha) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} L(\lambda).e^{(t+\alpha)}\, d\lambda$

The contribution to the error arising for any particular α is given:-

$$\text{Reducible error} = \int_{-\infty}^{o} |1(t+\alpha)|^2 \, dt: \quad \text{by change of variable:-}$$

$$= \int_{-\infty}^{\alpha} |1(t)|^2 \, dt: \quad \text{which is smallest when } \alpha \text{ is}$$

negative and large (i.e. a long delay). The total error is given by the sum of the irreducible and reducible errors. The knowledge of the magnitudes of the two components of error is a most important feature of the method and indicates whether a longer delay or an exact realisation of the network is warranted.

The result will be recognised as that derived by a more rigorous and arduous route by N. Wiener (4). The elegant and brief derivation given above was suggested by an unpublished note of the late Prof. P.J. Daniell (5) where he comments that by using the least squares criterion "no attention is paid to the advantages of a network which gives an exact reproduction as soon as possible. There is here a difficult problem to be though out which is to find the proper connection between the methods here and the practical criteria used at present in the design of predictors and filters. There is no reason to expect this problem to be insoluble".

TIME-WEIGHTED CRITERION

It is appropriate here to consider the possibility and merits of a criterion having some degree of time-weighting. It is well recognised that there are considerable difficulties in applying criteria other than least square which has particular simplicity. Little work in fact has

been done on any other basis. Dr. Shannon quotes several possibilities of evaluation functions in his paper on the continuous communication channel (6) but of these only least squares has any degree of practical application. Unfortunately there are many practical cases when least squares is not the most appropriate basis. This is the case when power amplifiers are concerned, in which the energy storage is of significance or when a limited power is available. In these cases, as in many others, by virtue of the subsequent usage of the information, there is particular merit in a criterion having a degree of time weighting. It remains to see whether an analysis on this basis is amenable to manipulation and whether the result is worth the candle. One thing is certain. The least squares criterion threw away the phase relations of the component vibrations. A time-weighted criterion cannot do this and consequently must retain them and be that much more complicated.

At first sight it might seem difficult to see how with a continuous, virtually unending, fluctuation it is possible to provide time-weighting. It is useful here to return to the transient description of the fluctuation. The transient may be treated individually and the mean error assessed finally. To each transient the network provides an output that is in error to the desired operation on it. The square of this error may be weighted according to the time since the time origin of the transient; this quantity integrated is finite. The mean value of errors to the complete fluctuation weighted in this fashion will also be a finite quantity and is the quantity to be minimised.

Let $e(t)$ be the difference between the total input and the output of the network and $E(j\omega).\overline{E(j\omega)}$ the spectrum intensity. Thus $E(j\omega)$ is the Fourier spectrum of the representative transient $e_r(t)$.

By Laplace transform theory:-

$$L\left\{e_r^2(t)\right\} = \int_0^\infty e_r^2(t)\, e^{-pt}\, dt = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} E(\lambda).E(p-\lambda)\,d\lambda$$

$$L\left\{t.e_r^2(t)\right\} = -\frac{\partial}{\partial p} L\left\{e_r^2(t)\right\}$$

$$L\left\{\int. t.e_r^2(t)\right\} = -\frac{1}{p}\frac{\partial}{\partial p} L\left\{e_r^2(t)\right\} \quad \text{and by the final value theorem:-}$$

$$M = L\left\{\int_0^\infty t\, e_r^2(t)\, dt\right\} = -\left[\frac{\partial}{\partial p} L\left\{e_r^2(t)\right\}\right]_{\lim p \to 0}^{\lim p \to \infty}$$

$$= -\lim_{p \to 0} \frac{\partial}{\partial p}\frac{1}{2\pi j}\cdot\int_{-j\infty}^{j\infty} E(\lambda).E(p-\lambda)\, d\lambda$$

since $\lim p \to \infty$ is known to be zero. Performing the differentiation and taking the limit:-

$$M = -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} E(\lambda).E'(-\lambda)\, d\lambda \quad \text{where } E'(\lambda) = \frac{d}{d\lambda}\left[E(\lambda)\right]$$

By change of variable $\lambda = j\omega$.

$$M = \frac{1}{2\pi j}\int_{-\infty}^{\infty} E(j\omega).\frac{d}{d\omega}\left\{\overline{E(j\omega)}\right\}\, d\omega$$

which is the quantity expressed in terms of spectrum intensity that has to be minimised. Using the previous notation for message and noise spectra

the quantity to be minimised in the case of discrimination is

$$2\pi j\, M_{\cdot} = \int_{-\infty}^{\infty} \left[G_{\cdot} e^{j\alpha\omega} - H.G_{\cdot} - H.M_{\cdot}\right] \cdot \frac{d}{d\omega}\left[\bar{G}_{\cdot} e^{-j\alpha\omega} - \ldots\right] \cdot \ldots$$

and by the calculus of variation this is a minimum when:-

$$\left\{ \bar{G}_{\cdot} \frac{d}{d\omega}\left(e^{j\alpha\omega}.G\right) - H.' F.\bar{F} - H.(G'\bar{G} + N'\bar{N}) \right\} = 0$$

$\Leftarrow$ bounded lower half ω plane

which expresses the condition to be satisfied by $H(j\omega)$ a stable network. An explicit solution has not in these case been derived; it will be appreciated however that the work is more difficult than on a least squares basis.

This is an appropriate point to summarise. The point is made, that the value and importance of a criterion is tempered by the difficulty of its application as well as by the quality of its achievement when applied. This might well serve as the keynote to discussion. An attempt has been made to show how the determinist philosophy is applied to the problem of prediction and in what manner it fails. The operational calculus of Heaviside and others achieves its success by virtue of transforming a discontinuous time function into a continuous function of a complex variable (although Heaviside himself would not have admitted the necessity for a complex variable).

By thus acknowledging a time origin it serves to qualify the determinist conception of prediction and use has been made of the calculus in deriving the best stable physically realisable operator for prediction of a transient.

An important extension of the application of the calculus occurs with its use in conjunction with statistical concepts in the analysis of populations of signals. The statistical aspects have been passed over in a rather superficial manner in order to emphasise the similarity between the two methods of application of the calculus. Possibly this obscures the originality of the departure from determinist concepts which is essential if the calculus is to play a part in the theory of information. More subtle criteria of the significance of waveforms than have been entertained here will be related to the statistical analysis of the waveform. Meanwhile it has been demonstrated that a time-weighted criterion has possibilities.

REFERENCES

1. James H.M., Nichols, N.B. Phillips, R.S. "Theory of Servomechanisms" Radiation Laboratory Series Vol. 25 p.61 McGraw-Hill, (1947).

2. Lee, Y.W. and Stutt, C.A. "Statistical Prediction of Noise". Technical Report: Research Laboratory of Electronics, August (1949).

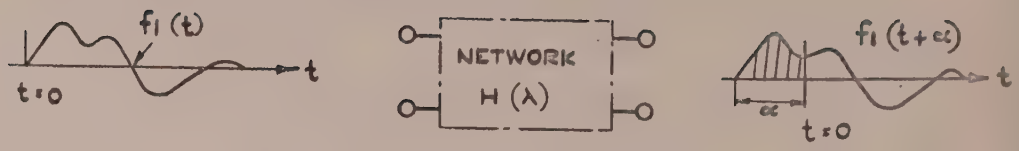
3. Hopf, E. "Engoden theorie". Erg. d. Math. (1936).

4. Wiener, N. "Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications". Wiley, (1949).

5. Daniell, P.J. "Digest of Manual on Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications by Norbert Wiener". SRID Progress Report 328.

6. Shannon, C.E. "A Mathematical Theory of Communication" BSTJ. October, (1948).

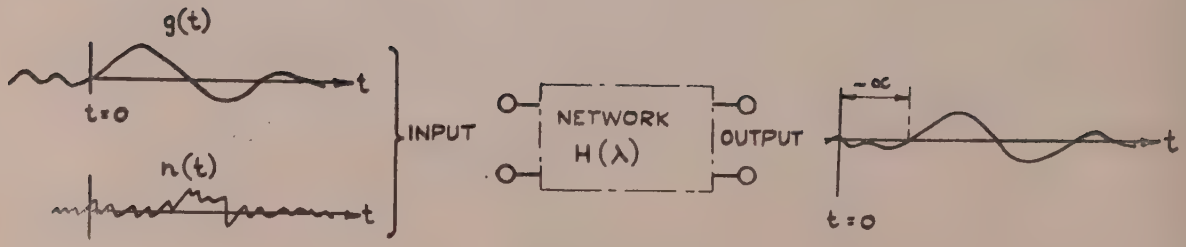
FIG. 1.



INPUT SIGNAL.

OUTPUT SIGNAL PREDICTING α SECS. AHEAD.

FIG. 2.



ENTROPY, TIME AND INFORMATION
(INTRODUCTION TO DISCUSSION)

by
D.M. MacKay

(1) INTRODUCTORY

To 'stray outside one's field' has become no longer the innocent sign of an enquiring mind which it was in the time of Newton or Boyle. In a gathering as catholic in its interests as this one, however, it may be less reprehensible to seek to exchange ideas on matters which are nobody's private preserve. The present paper is intended only as the opening contribution to such a discussion, on one of the older problems in the background of Physics. Do the concepts and the approach of Information Theory help to shed light on what we mean by Time, and the nature of the Second Law of Thermodynamics? The notes which follow are much condensed, but it is hoped that they may stimulate some constructive answers to this question.

(2) ENTROPY AND INFORMATION

The relation between the selective information content of a signal and the mathematical notion of entropy has been discussed in detail by Shannon, Weaver, Gabor and others [1,2,8]. It has been pointed out that it should not facilely be equated with thermodynamic entropy. The connexion, in one common and typical case, will now be examined. It will serve our purpose to consider a single measurement, of voltage V across resistance R at temperature T, which is assumed to occupy the minimal time Δt appropriate to a bandwidth Δf. Let us assume that only thermal noise limits precision.

The result will be a value of V which can be represented by the number of intervals n which it occupies on a dimensionless proper-scale (q.v.) calibrated in steps proportional to noise voltage. The amount of metrical information or metron-content i of the result is the square of n, and is thus proportional to the signal:noise power ratio. Without repeating detailed calculations [4] here, we may note that this is inversely proportional to T, and directly to the product $V^2 \Delta t/R$. It is in fact approximatly $(1/k)$ times the minimum increase in entropy $(V^2 \Delta t/RT)$ which must occur during the measurement. (cf. Szilard, Ref. 9.)

Thus in this typical case one unit of metrical information costs the system about k units of entropy-increase at least. What of the selective information-content? This is measured, in the case of a signal, as the logarithm of the number of equiprobable states. The assumption is usually made that all quantized values of voltage up to a certain limit are equally probable, for each logon. (The generalisation to the multidimentional case makes no difference in principle.) Thus our single logon provides $\log_2 n$ bits of selective information. This is in fact the entropy of the selection made by the signal.

Evidently n^2 units of (dimensionless) thermodynamic entropy here yield not more than $\log_2 n$ bits of "selective entropy". There is in fact no logical connexion between the two, because our assumption of equiprobability - (i.e.) of a rectangular probability-distribution for V - represented our knowledge not of the statistical behaviour of the apparatus as a physical system at temperature T, but of the statistical behaviour of a distant sender. In other words, our ensemble of possible states has not at all the same composition as the ensemble appropriate to a physical system in thermodynamic equilibrium.

It is for such an ensemble that the statistical definition of physical entropy holds. And since the probability of a logon's having the energy-level E is classically proportional to $e^{-E/kT}$ at equilibrium, the negative logarithm of this is linearly related to E/kT and hence to n^2 or i, the metron-content. In fact the definition of entropy-change as $\Delta E/T$ is equivalent to the statistical definition only on this assumption.

We find therefore that the metron-content of a measurement is linearly related to the statistically-defined entropy of the system when this is calculated for the ensemble representing thermodynamic equilibrium. In short, to use Fisher's term[3], our ensemble is based on the null-hypothesis: the hypothesis that no signal is present.

It will be remembered[4] that we consider metrons to signify the occurrence of elementary events. If each of these has a prior probability p of being associated with a given logon, the prior probability of i being so associated is p^i. The logarithm of this is again proportional to i.

Finally it may be recollected that Fisher himself has remarked on the general analogy between his definition of information (see Glossary) and the notion of entropy. In our case his measure would be (1/kT) times the "size of sample". Although this measure requires normalization before comparison with others, it is evident that the "size of sample" is proportional here to energy. Thus Fisher's definition of information, like our metron-content, is linearly related to the selective information-content or entropy of the selection made from an ensemble prepared on the null-hypothesis.

We shall return to this point after considering the notion of time.

(3) TIME

In discussing time it is easy to introduce the concept in a circular way, via adjectives such as "later" or "next". To avoid this let us suppose that we begin with a set of "states of awareness" or "representations of the case", in disorder. What do we mean by saying that these can be set in a time-order, and how could we accomplish this?

Briefly, it is suggested that the relation "later than" between states is always correlated with a relation "greater than" between the corresponding values of some information - measure of representations of those states. The concept of time however has several connotations. These appear to arise, as one might expect, through implicit reference to different kinds of information-pattern. To summarize:

(a) We are subjectively aware of the relation "later than" between states of awareness; a "later" state is always (and by definition?) one in which our information-pattern representing observed events is larger - or rather, has had an addition. (The test of time-order is here of course a differential one, and questions of memory-failure are irrelevant.)

(b) We find that certain numerical features of the physical world (distance traversed by the tip of a "constantly" rotating vector for example) show the same order. (By "constantly" we appear to have in mind a negative criterion. Uniform motion is motion about which nothing singular can be said. A system capable of perfectly uniform motion is ideally a one-parameter system. A single elementary proposition says all there is to say about it in the relevant respect. Hence as standards of objective time we choose systems (oscillators, rotators) as near to this ideal as possible.)

Distance is however a vector quantity, and this concept of time is not uniquely directed. Indeed, as we have seen elsewhere (6), it merely provides a co-ordinate-framework, which extends to both infinities as the logical representative of the single parameter of frequency. For the "arrow" of objective time we must look elsewhere. It may be noted however that time-interval in this sense is measurable in terms of the dimensionality of the information-space in which we are able to describe our total scientific (and hence presumably objectively-based) awareness.

(c) Mere ticking of a clock however does not convey the passage of time. Subjectively each tick correlates with a particular subjective information-pattern in a growing sequence, and so acquires a subjectively assigned direction. But it seems reasonable to expect our concept of the objective passage of time to be derived, like other abstract concepts, by analogy from subjective experience. Can we then find an analogous information measure in the objective world, to give time its arrow?

(4) THE NON-CREATION OF INFORMATION

There is a kind of folk-saying in Physics, which has survived even the onslaught of Quantum Theory. It asserts that there is "no effect without a cause". It is perhaps arguable that it is an axiom or even a tautology when suitably expressed. It is illustrable by the unfailing success with which we can sort sufficiently complete photographs of a diffusion-process or an explosion, into the "correct" time-order. The miraculous, in other words, is not normally expected in science.

It is suggested that this principle can be translated as: "No information without a source", or: "Information is not normally created". This means in practice that when our representation of what is the case changes, we say we have received information from the system concerned. This suggests that it has lost something; and in fact we are now able to give this quantitative meaning. When we look into the matter we have found that the entropy of a system must increase at least in proportion to the metrical information we gain from it. Normally and indeed essentially our coupling to a system is so inefficient that we gain only a small fraction of the available information; but in principle it appears that the states of isolated physically-knowable systems (i.e., isolated from information-sources) must have the same order on an entropy-scale as their representations do on a scale of metron-content. In other words, when we set out representations of states of awareness in order of increasing metron-content, the corresponding states themselves are necessarily in order of increasing entropy.

(5) TIME AND THE SECOND LAW OF THERMODYNAMICS

It seems then to be a tenable hypothesis that in every context the concept of time can be identified with a measurable property of the corresponding representation. Is this what we mean by time? It is difficult to think of a context in which it could not apply.

Our basic suggestion has been that the intuitively-known relation "later than" is synonymous with the relation "greater in information-content than", referring to an information-pattern. We have seen this to accord with the subjective concept of time, and it has been suggested that our objective notion finds meaning only when it relates to a suitable analogous information-pattern. We then find that if we accept the principle of non-creation of information, all knowable physical systems isolated from systems giving out information, possess greater entropy the greater the information they have given out. If then order on our information-scale means order on our time-scale, is the second law of thermodynamics a tautology?

There are other implications of this hypothesis, in cosmological models, for example; and the question of the magnitude of time-intervals has not here been raised. But perhaps this question is enough for one discussion.

Two remarks by Weyl[7] may perhaps be quoted in conclusion:

"The idea inherent in causality, that that which is earlier is the determining reason for what follows, and vice versa, impresses on our probability judgments a distinguished direction in time".

"Our judgment thus proceeds as if the system with which we are dealing had been created before our time. The word "creation" suggests a metaphysical or even theological interpretation, but this should not prevent us from recognising the state of affairs which is most aptly expressed by this word".

Has metaphysics caught up with us after all?

REFERENCES

1. Shannon C.E. & Weaver W. "Mathematical Theory of Communication": U. of Illionois Press (1949).

2. Gabor, D. "Communication Theory and Physics": Symposium Paper (see page 48)

3. Fisher, R.A. "The Design of Experiments": Oliver and Boyd (1935)

4. MacKay, D.M. "Quantal Aspects of Scientific Information:" Symposium Paper.

5. Fisher, R.A. J.Roy.Statist.Soc. 98, 39, (1935)

6. Mackay, D.M. Phil. Mag. Ser.7 41, 302

7. Weyl, H. "Philosophy of Mathematics and Natural Science": p. 204 Princeton (1949)

8. Wiener, N. "Cybernetics": Chapters II & III Wiley & Sons (1948)

9. Szilard, L. "The Diminution of Entropy in a Thermodynamic System caused by the Intervention of Intelligent Beings": z. fur Physik, 53, 840 (1929)

DISCUSSION

The following contributions to the discussion are those only which have been communicated in writing. No record is included of the general discussion which took place at the Meeting.

These contributions have been arranged in the same order as that in which the various papers were delivered.

DISCUSSION ON MR. E.C. CHERRY'S PAPER "A HISTORY OF THE THEORY OF INFORMATION".

PROF. B van der 'POL.

(1) In connection with the language question raised, the following may be of interest. In the International Telecommunication Union extensive use is made of what is called "a simultaneous translation system", where a translation takes place during the actual speech of the speaker. I asked the head of the interpretation service how long was the average delay between the original speech and the translation. The answer was that this delay depended principally upon the language from which the translation took place rather than the language into which it was being translated. If the original language was English this delay was of the order of 5 to 7 seconds, whereas if the original language was German this delay was usually much longer, up to 15 to 20 seconds or so.

(2) Some time ago in Holland an enquiry was made asking whether, when any new thought in science occurred to one, this new idea occurred in words or not. The question was also addressed to me. My answer was positively no, because it often gave me considerable pains, after a new thought had occurred to me, to express it in words to friends.

MR. P.M. WOODWARD.

I think it a little unfortunate that whenever the uninitiated come across the principle of inverse probability, they find it associated with such phrases as "thorny question", "inapplicable", "general disagreement". It should be made quite clear, as Mr. Cherry has in fact done, that when a priori probabilities really exist, as they often do exist, the principle is perfectly straightforward and follows directly from the elementary laws of probability. Suppose, for example, it rains on four days for every three it is fine, that when it rains the glass is low three times out of four, and when it is fine it is high two times in three. We thus have the following sample:

Rain	Rain	Rain	Rain	Fine	Fine	Fine
Low	Low	Low	High	High	High	Low

Now suppose we have access to the glass but not to the weather. There are two hypothesis, Rain and Fine, and there is some data, High say. Then the probability that it is in fact Fine is two thirds. That is all Bayes' Theorem states. The axiom is not used, and in a problem of this type, there is no disagreement at all. I think it important that we have this clear at the outset because Dr. Shannon's theory does involve the principle implicity.

DR. I.J. GOOD.

There are three comments which I should like to make arising out of Mr. Cherry's interesting lecture though I fear that these remarks have rather high redundancy. The first is concerned with what is meant by a 'reasoning machine'. The phrase 'reasoning machine' has at least six different possible interpretations which can be exhibited in tabular form. (See diagram.)

	DETERMINISTIC	INDETERMINISTIC
Formal	1	2
+ probability	3	4
+ aesthetics	5	6

The numbers 1 to 6 give the chronological order in which the machines will probably be built. When I talk about a reasoning machine I tend to mean one of type 3 or 4. This usage is consistent with the definition reasoning = logic + probability. No practicing statistician is a machine of type 1. The behaviour of any of the machines 2, 4 or 6 is not completely predictable by the programmers (one possibly hardly predictable at all). Machines of type 2 can be used in order to obtain empirical estimates of rounding-off errors in computations, by doing random rounding-off. They also have other more sophisticated uses such as the so-called Monte-Carlo method of solving ordinary mathematical problems. A machine of type 5 may eventually be built, since N. Rashevsky of Chicago has had a little success in reducing aesthetic judgments to rule of thumb.

Notice that the classification is only a functional one. For example, from a practical point of view types 2, 4 and 6 can be included in types 1, 3 and 5 by merely storing random numbers.

My second comment is concerned with the first equation on p.25 of Mr. Cherry's paper.

Shannon is concerned with only one set of input probabilities $p(\mathcal{H}_1)$, $p(\mathcal{H}_2)$, ..., σ, as we may say, with a known input language $\mathcal{H}$. It is interesting to consider the problem of trying to determine which of two alternative languages occurs at the input, in terms of what is observed at the output. This is distinct from the problem of determining in input message. The relevant equation is:

$$\log O(\mathcal{H}|E_i) - \log O(\mathcal{H})$$
$$= \log P(E_i|\mathcal{H}) - \log P(E_i|\overline{\mathcal{H}})$$
$$= \text{weight of evidence in favour of } \mathcal{H} \text{ in virtue of } E_i.$$

Here E_1, E_2, ..., E_N are supposed to be the mutually exclusive possible outcomes of an experiment; O means odds = $p/(1-p)$, where p is a probability and $\overline{\mathcal{H}}$ means 'not $\mathcal{H}$'. (MacKay uses the phase 'weight of evidence' in a different sense). The equation is a simple deductions from Bayes' theorem. If $-\log P(E_i|\mathcal{H})$ is called 'the amount of information from S_i, assuming $\mathcal{H}$', we see that the weight of evidence is in favour of the hypothesis which gives less information.

Finally I have a query concerning the number of functional effector and receptor units in the body. (See the notes, p. 22, l. 13). If there are oriented connections between every pair of the 1000 units, the number of connections would be only 10^6. How is the figure of 10^9 obtained?

MR. CHERRY IN REPLY:

Since my paper has been intended as a historical survey of our subject, I have taken the liberty of re-drafting certain sections, taking advantage of the points raised by the various speakers

DISCUSSION ON DR. SHANNON'S PAPERS.

MR. E.C. CHERRY.

There is a well-known elementary way of interpreting the "selective entropy" expression for the information conveyed by a symbol-sequence, which serves as an introduction to the subject, and which should perhaps be recorded. Consider one symbol, having a known probability of occurrence Pi, in a code of n such symbols. It is reasonable to assume that the "information" conveyed by this one symbol is the least number of selections, H, needed to identify it amongst the n in the code. Arrange the symbols in order of decreasing probability P_1 P_2 .... P_2 .... P_i (total probability = 1.0); divide into two groups (P_1 P_2 .....) and (... P_i ... P_n) of equal total probability $\frac{1}{2}$; again divide the group containing P_i into two, of probabilities $\frac{1}{4}$. Continue such bisection H times until two groups remain, each of probability P_i , one being the wanted symbol. Then:

$P_i \cdot 2^H$ = total probability of the symbols in the code = 1.0
or $H = -\text{Log}_2 P_i$ (1)

The average number of selections required for a complete message is then the mean of H or

$$H_{\text{average}} = -\sum_i P_i \log P_i \qquad (2)$$

This argument assumes of course that the symbols may always be divided into two groups of equal probability; it perhaps has the merit of emphasising the reasonable nature of the expression (2) as representing information.

MR. S.H. MOSS

During the discussion following Dr. Shannon's second talk, Professor Van Der Pol raised the question of what is meant by the delay imposed on a transient waveform by a process which at the same time distorts it.

If the process is linear, and has a finite zero-frequency response, the time lag between the (temporal) centroid of the output transient and the centroid of the input transient is a constant, which is a characteristic only of the system, and is independent of the wave-form of the input transient. It is thus an appropriate measure of delay. Its value is the slope of the phase-shift versus frequency curve at zero frequency.

For a wave-packet, considered as a sinusoidal wave of reference frequency, modulated in amplitude and phase by a transient complex envelope, there is an acceptable sense in which the centroid of the envelope is delayed by a constant time interval, independent of its waveform, if the amplitude versus frequency characteristic of the system is finite and stationary at the reference frequency. Here again, its value is the slope of the phase-shift curve at the reference frequency, the well-known expression for the group-delay. In the general case, when the amplitude characteristic of the process is not stationary at the reference frequency, the situation is more complex.

Each of these results is a special case of a class of additive invariants associated with linear systems. They are closely analogous to the cumulant statistics used to described the properties of a univariate statistical distribution and of its characteristic function (i.e. its Fourier transform) in sampling theory.

Dr. Uttley

Concerning the mistakes made by an automatic computer it is the principle of redundancy which can contribute to a solution.

Firstly one can incorporate redundant equipment, checking circuits for example, and in the limit by employing two computers as mentioned by Prof. A.V. Hill. At present, however, the designers of large computers quite reasonably are loth to take this step. As a result present machines possess the property of a nonredundant code that a single error or change produces a quite different result; this is intolerable.

Redundancy can be incorporated in a second way. When a number is fed into a machine, additional redundant digits can be introduced with it; their function can be to indicate the presence and location of errors in the number. This redundancy can be obtained at the expense of speed of operation of the computer.

Dr. Shannon pointed out that the specialized theory of coding called by him "time reserving theory" is far more important with practical aims in mind. But would he not agree that from this practical point of view, it should be still better to deal with an unfortunately much more difficult case - I mean the case of a given definite time of coding operation?

Dr. I.J. Good.

I would like to mention very briefly a mathematical curiosity which may be of some significance.

Consider a source of information which produces digits of N types with independent probabilities $p_0, p_1, \ldots, p_{N-1}$. Imagine an infinite sequence of such digits produced and prefixed by a decimal point (or rather an N-imal point). Then the resulting point will almost certainly belong to a set of points of Hausdorff-Besicovitch fractional dimensional number equal to the relative entropy of the source.

Mr. W. Lawrence

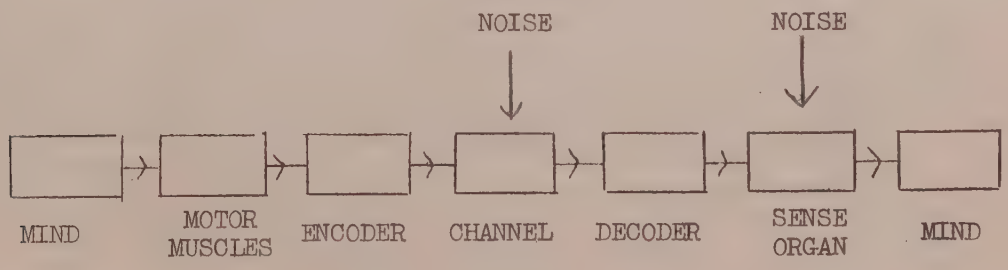


FIG. 1

In consideration of the block schematic of Fig. 1, it has commonly been assumed that the only function of the decoder was to restore the message to the form presented to the encoder, in order that the message might be "understood". Alternatively, the message might be restored to some other understandable form, as when a message originally spoken is transmitted as a telegram and presented to the receiving mind in writing. In either case the decoder operates to increase the redundancy of the message, and it is this increase in redundancy that I wish to talk about.

The mind can only accept as information, material that is presented to the senses with a considerable degree of redundancy.

Random acoustical noise, or random scintillations on a television receiver mean nothing. The more highly redundant the material presented to the senses, the more effortlessly does the mind receive it, provided of course, that the redundancy conforms to an agreed convention that the mind has been educated to accept.

We can just apprehend speech presented with a 1000 c.p.s. band width and a 15 db signal noise ratio, though to do so requires considerable mental effort. Speech with a band width of 3000 c.p.s. and a 40 db noise ratio can be apprehended without appreciable conscious mental effort. With a band width of 15,000 c.p.s. and a noise ratio of 60 db we feel a marked improvement which is especially appreciated when we are listening to something difficult to understand, such as a philosophical lecture.

We can express the channel capacity required for these presentations in bits/sec. by considering the P.C.M. channel that will just handle them.

The inherent information rate of spoken English is, say, 100 bits/sec. The channel capacities required for the three presentations considered above are roughly 5000 bits/sec., 50,000 bits/sec. and 500,000 bits/sec., representing "minimum tolerable", "good commercial" and "near perfect" presentations.

It is also interesting to consider telegraphy, presented to the senses and the mind as written matter. The channel capacity required for the material presented to the eyes of the recipient can be assessed by considering a P.C.M. television channel just adequate for the presentation considered. The number of digits in the Pulse Code is controlled by the extent to which the blacks and whites of the writing stand out from the random specularity of the background. The number of elements in the picture is controlled by faithfulness of the reproduction of the letter forms and the number of letters or words simultaneously visible. The number of frames per second is controlled by the desired steadiness of the picture.

A "minimum tolerable" presentation might be 3 digit P.C.M., 50 elements per letter, 5 letters simultaneously visible and 10 frames per second, which requires a channel capacity of 7500 bits/sec. A "good commercial" presentation, as good as a ticker tape, requires a channel of about 10^5 bits/sec. and a "near perfect" presentation such as first class printing with a whole page simultaneously visible, requires about 10^8 bits/sec. This again is the condition we like to get when trying to understand something really difficult.

The higher channel capacities requires for the written presentation are consistent with the fact that we can read language faster than we can listen to it, and also to the fact that we prefer a written presentation when the subject matter is really difficult.

I believe that the habit of only attending to redundant material is a defence mechanism that the mind adopts to sort out information worth attending to, from the inconceivably vast volume of information with which the senses continually bombard it.

This also clears up a paradox that used to worry me and may have worried others. Instinctively we feel that a "Random" sequence contains no information, whereas an orderly sequence "means something". Communication Theory, however, says that a random sequence contains maximum information and that a completely ordered pattern contains no information at all. I would explain this by saying that the more nearly a sequence is random the harder it is for the mind to comprehend, and in the limit it contains maximum information which is, however, totally incomprehensible.

Even a machine requires some redundancy in the signal, such as the synchronisation digit in P.C.M., before it can "comprehend" or "decode" it. It can work with very little because it knows just what to look for, and its attention is not distracted. Our mind and senses, which have been evolved in a highly competitive environment, demand much more.

Mr. W.P. Anderson.

1. The proposal to investigate a system in which the interference takes the form of a random binary signal added to the wanted signal is an interesting one but such a system is a very long way from the noisy communication channel with which the engineer is concerned. The errors in such a system are absolute, as the errors are assumed to be in the error correcting code described in the paper, that is to say a single element of the code is received either correctly or incorrectly.

In a physical communication system however these are no absolute errors. From the amplitude of the received signal element the probability that the transmitted element was a "mark" can be computed. If the signal to noise ratio is large, this probability will almost always either be nearly unity or nearly zero, depending on whether the transmitted element was in fact a "mark" or a space", but if the signal to noise ratio is small it may have any value. This probability contains all the information obtained as a result of the reception of the element and if the system at some stage distinguishes between only two classes of elements, those having the greater probability of being "mark" and those having the lesser probability of being "mark", information is being discarded. Error detecting codes necessarily operated in this way hence it would appear that they must be less effective than integrating systems in which the amplitudes of individual elements are preserved.

This conclusion is of some interest, apart from its application to error detecting codes, as integration is equivalent to narrowing the band and it suggests that no advantage is to be gained by increasing the baud speed of a telegraph transmission and introducing a code containing more than the minimum number of elements per character. It is believed that this conclusion is correct for double current working where the energy required to transmit a character is simply proportional to its length, but not for single current working, where the energy required to transmit a character of given length varies over a wide range. In the latter case increasing the speed increases the number of possible characters and the number actually required can be selected from those requiring least power to transmit. In the limit of course as the bandwidth is increased such a system reduces to Pulse Position Modulation, with one mark element per character.

2. A somewhat similar loss of information arises in the process of "quantisation" in pulse code modulation systems and it would appear that it must always be better in principle to send the residual amplitude or "error signal" instead of the least significant digit.

3. Information theory has so far dealt with signals and noise superimposed in linear systems. In a system of great practical importance however, long distance radio communication involving ionospheric propagation, the transmission path itself fluctuates in a manner which is only definable in a statistical sense. It would appear that the existence of such fluctuations must reduce the rate at which information can be passed over the link and that the extent of the reduction should be determinable by the methods of Information Theory. It is hoped that some attention will be given to this problem in the further development of the theory of information.

Dr. C.E. Shannon. (In reply).

The point raised by Dr. Uttley regarding the use of redundancy for error detection and error correction in computing machines is certainly an important one, and will become more so as the machines become larger and more complex. In addition to the two methods pointed out by Dr. Uttley, redundancy in equipment and redundancy in encoding, a third may be added, redundancy in programming (as, for example, in redoing the calculation a second time on the same computer). The first two methods require additional equipment and the last additional time.

The present Bell Laboratories Relay Computer as well as some of the previous designs use both of the first two methods, the third of course being optional in any computer. All the relays are furnished with twin contacts. In effect, this amounts to a second at least partially independent computer, paralleled with the first one at the most critical points, the relay contacts. Furthermore, the numbers are represented in a two-out-of-five code; each decimal digit is represented by a group of five relays, and the code is such that exactly two of the relays must be operated to represent a digit. If this check fails at any stage the machine is automatically stopped. The circuit is such that any single error will be detected.

This is an example of an error-detecting scheme, which works exceptionally well. If errors were more frequent, it might be advisable to introduce an error-correction system in such a way that any single error would be corrected automatically by the machine, while two simultaneous errors would be detected and cause the machine to stop. It is possible to encode a decimal digit into seven binary digits and obtain single error correction. With eight binary digits, a code can be found which gives single error correction and double error detection.

Concerning the points brought up by Mr. Anderson, the error-correcting system suggested in the paper was meant particularly for applications such as computing machines where the information is encoded into a binary system with a definite reading of zero or one. In a pulse communication system with additive Gaussian noise a preliminary integration process followed by a threshold device gives a binary indication of whether the pulse was there or not, and in fact it can be shown that by proper choice of the weighting function in the integration such a detection system divides all possible received signals properly into two classes, those for which the a posteriori probability is in favour of a pulse and those for which it is not. Thus such a detection system is ideal in such a case in the sense of making the fewest possible errors for individual pulses. However, if this error frequency is still too high it may be desirable to introduce redundant encoding and error correction.

Information theory has by no means been limited to linear systems, although some of the special results apply only in these cases. Statistical variations in path length, etc., must of course be considered a form of perturbing noise, and the channel capacity and proper encoding systems, can, in principle, be calculated from the usual expressions, although such calculations are, because of their complexity, usually impractical.

M. Indjoudjian has raised the question of what might be called a finite delay theory of information. Such a theory would indeed be of great practical importance, but the mathematical difficulties are quite formidable. The class of coding operations with a delay $\leq T$ is not closed in the mathematical sense, for if two such operations or transducers are used in sequence the overall delay may be as much as $2T$. Thus we lose the important group theoretic property of closure which is so useful in the "infinite delay" and "time-preserving" theories.

Nevertheless, any results in a finite delay theory would be highly interesting, even if they were restricted to the solution of a few special cases. Some work along this line appears in a recent paper by S. O. Rice (Bell System Technical Journal, Vol. 29, January 1950, pp. 60-93), where estimates are made of the probability of errors with various delays when attempting to transmit binary digits through white thermal noise.

Mr. Lawrence has pointed out that the brain can generally accept information only in a highly redundant form. It seems likely that the reason for this lies in the fact that our environments present us with highly redundant information. The scenes we view are well organised and change relatively slowly, and the significant sounds we hear tend to be localized in pitch and to persist much longer than this localization required. Nature, then, would design the nervous system in such a way as to be an efficient receptor for this type of information and to make use of the redundancy to achieve higher resolving power and better discrimination against noise. Experiments in psychological optics have, indeed, shown that the eye can determine if two line segments lie in a straight line much more closely than the width of a rod or cone, or of the diffraction pattern of the lines in question, thus showing that the eye makes use of this redundancy to improve discrimination

The number of nerve cells in the optic nerve is only about one per cent of the number of rods and cones in the retina. If the time constants of both elements are about the same, this implies that the capacity of the optic nerve for transmitting information to the brain can be only about one per cent of the information that would be received by the retina. Thus only if this information is highly redundant could it all be encoded into a signal to be transmitted via the optic nerve to the occipital lobe. At that point further abstraction of the basic information, i.e., elimination of redundancy, probably occurs in the connections with the related association areas.

Mr. Good has pointed out an interesting relation which I had also noticed between entropy and the Hausdorff-Besicovitch dimensions number. While it is easy to see the reason for this from the basic definition of Hausdorff-Besicovitch dimension number and certain properties of entropy, I believe the root of the relation springs from the following consideration. A dimension number to be reasonable should have the property that it is additive for product-spaces, that is, the set of ordered pairs (λ,μ) should have dimension number $d_1 + d_2$, where d_1 is the dimension number of the set (λ) and d_2 that for (λ). Similarly, a measure of information should be additive when we combine two independent information sources, i.e., a stochastic process producing ordered pairs, one from each of two independent sources. These desiderata result in the logarithmic measures which appear in both fields.

DISCUSSION ON DR. D. GABOR'S PAPER "COMMUNICATION THEORY AND PHYSICS".

PROF. B. VAN DER POL.

The expression synthesizing a function f(t) (which contains no frequency components higher than critical angular frequency ω_o) from its equally spaced discrete values $f(\frac{\pi n}{\omega_o})$, where n = ...-2,-1,0,1,2..... is

$$f(t) = \sum_{-\infty}^{+\infty} f(\frac{\pi n}{\omega_o}) . \quad \frac{\sin(\omega_o t - \pi n)}{\omega_o t - \pi n}$$

The individual contributions

$$C_n \frac{\sin(\omega_o t - \pi n)}{\omega_o t - \pi n} = C_n \frac{\sin \chi}{\chi} \quad , \text{ say,}$$

where $$C_n = f(\frac{\pi n}{\omega_o})$$

and $$\chi = \omega_o t - \pi n$$

fall off as $1/\chi$ and are therefore localised to a small degree only, because $\frac{\sin \chi}{\chi} = O(\frac{1}{\chi})$ as $\chi \to \infty$.

It is possible to localise the individual contributions considerably more by the following procedure. We note that

$$\frac{\sin \chi}{\chi} = \sqrt{\frac{\pi}{2}} \frac{J_{\frac{1}{2}}(\chi)}{\chi^{\frac{1}{2}}} = O(\frac{1}{\chi})$$

and the general relation

$$\sqrt{\frac{\pi}{2}} \frac{2^n J_{n+\frac{1}{2}}(\chi)}{\chi^{n+\frac{1}{2}}} = \frac{1}{\pi n} (1+\frac{d^2}{d\chi^2})^n (\frac{\sin \chi}{\chi}) = O(\frac{1}{\chi^{n+1}}).$$

Hence, if instead of considering the function f(t) itself, we construct the function

$$(1 + \frac{d^2}{d(\omega_o t)^2})^n . f(t)$$

we see that the individual contributions in the series expression for the above function are considerably more localised round the points $f(\frac{\pi n}{\omega_o})$ than is the case for the function f(t) itself.

DR. J. J. GOOD.

I cannot help wondering whether it is not largely a prejudice to analyse signals in terms of frequency.

The prejudice arises partly because of the desire to ignore the apparatus which is to be used for interpreting the signal, f(t). When the behaviour of the apparatus can be described by means of a differential equation of the form.

$$\frac{d^2 \chi}{dt^2} + \omega^2 \chi = f(t)$$

it is natural to think in terms of sine waves and frequencies. But this analysis is less appropriate when, for example, there is a term in $\frac{dx}{dt}$.

Similarly it may be that in quantum theory a lot of philosophical difficulties arise merely as a consequence of the insistance on analysing events in terms of frequency.

DR. D. K. C. MACDONALD

(1) The question of the ultimate "meaning" of the duality of frequency and time leads one to recall the very fundamental problem of the "meaning" of Planck's constant and its unit of measure - action. Although this may be expressed as energy x time or momentum x distance, etc., it still appears very difficult even yet for our brains to visualise although we have had Hamilton's Principle of Least Action available for so long as an analytic tool in dynamics.

(2) The importance of elementary $\frac{\sin t}{t}$ functions in this field (as in the "sampling theorem") calls to my mind some earlier essays in realisable network theory* where one finds that network having an amplitude-response $\sin f/f$ (and related derivatives) are inherently realisable. It appears to me now probable that the work could be extended to more general networks than seemed profitable at that time.

DR. D. GABOR (in reply)

Dr. Good has raised the interesting question whether the widespread use of Fourier analysis in communication problems has other grounds than just habit or prejudice? There are in fact two good reasons for this preference. In communication problems we deal usually with the infinite or semi-infinite time axis, moreover the problems are usually homogeneous in time. Once one or the other of these conditions is dropped, it may be well worth while to carry out the analysis in terms of other functions. For instance if the time interval considered is finite but homogeneous, gaussian elementary functions or functions of the type $\sin\omega t/\omega t$ may have advantages. If there is a distinguished instant of time, Hermite's orthogonal functions may be preferable. If both conditions are dropped one might choose Laguerre functions, Legendre functions Tchebysheff polynomials and a host of others. One can even argue that in communication problems in the narrower sense of the word there is always a distinguished instant, viz. the present, and in fact functions with a "perspectivic" view of the past may be the best.

Regarding the question posed by Dr. Macdonald, I do not think there is any problem in the ultimate "meaning" of the quality of frequency and time; they are simply dual by definition. The problem of the elementary action is quite another matter; it is one of the most fundamental of crude facts, and extremely refractory to visualization. It does not help much if one tries to "visualize" a momentum as the frequency conjugate to a co-ordinate. Dirac's remark is worth remembering: "The main object of physical science is not the provision of pictures but the formulation of laws governing phenomena".

* E.g. MacDonald, D. K. C. : Phil. Mag. 38, 115,(1947)

DISCUSSION ON MR. MACKAY'S PAPER "QUANTAL ASPECTS OF SCIENTIFIC INFORMATION"

DR. I. J. GOOD

Is "this cat is black" an atomic proposition, or is the idea of atomic propositions a purely abstract concept?

/Later./ Is "this cat is in contact with this box" an atomic proposition? On the face of it it should be since it is a coincidence relation.

PROF. M. S. BARTLETT

Could Mr. MacKay clarify the trend of his paper in its relation to the preceding paper by Dr. Gabor, and to Lord Cherwell's work, referred to by a previous speaker, on the interpretation of quantum theory? Whereas Dr. Gabor had started from quantum theory and deduced the ultimate 'bits' of communication theory from it, Mr. MacKay seemed to be starting from the other end, and attempting to deduce quanta from logical analysis into atomic propositions. Mr. MacKay would perhaps agree that this logical approach at best left an undetermined constant which could only be obtained by experiment.

MR. A. H. REEVES

If we accept the idea that it is allowable to quantise all measurable 'information' it may be of interest to note that the location of any point in a finite hyper-space may be defined by (a) a finite scalar quantity, or (b) a finite integral pure number. In case (a) we may suppose the hyper-space to be scanned by any line of fixed but arbitrary shape such that the maximum interval between any two adjacent points unscanned is equal to the minimum value that can be measured with a pre-determined precision-probability. The length of the scan line, from an arbitrary starting point, then gives the required scalar definition of the point's location.

In case (b) we suppose in addition that the scanning line itself is similarly quantised, attaching a different integral number to each quantised length along this line. Although it seems difficult to define quantising and scanning constants that would give useful results, this process would at least avoid the multiplicity of dimensions inherent in vector concepts.

As one example, it is clear that we could in this way combine the units 'logon' and 'metron' into a third single unit containing both. Whether or not this would be a useful thing to do is of course another question.

DR. P. A. MORAN

There seems to be a fundamental philosophical confusion in Mr. MacKay's paper. It is quite useful and interesting to set up a more general method of measuring information than that used by Dr. Shannon. But it must be remembered that to split up scientific statements into "atomic propositions" is a conventional and ad hoc procedure which, however useful for certain scientific purposes, is quite mistaken if taken as a serious description of epistemology. To confuse a scientific method or a scientific theory with a philosophical one is nowadays far too common a mistake.

MR. D. M. MACKAY (in reply):

I should perhaps begin by emphasising that in adopting the concept of an atomic proposition I am merely applying the technical results of

Wittgenstein and others to the field of physical science, for which I believe they hold good. I would in fact repudiate the positivist conclusions (or lack of conclusions) in other realms of thought which by their nature do not admit usefully of exhaustive analysis in terms of atomic propositions. But the appeal of physical science is to sequential (and therefore discreet) verification; and I know of none of the current objections to Wittgenstein's philosophy which would even claim to apply to the use of his propositional theory in physical science.

In reply to Dr. Good: an atomic proposition asserts an atomic fact. An atomic fact is a fact which has no parts that are facts. But the named things entering into meaningful sentences are seldom 'simples', since we define them (implicitly if not otherwise) by asserting other facts. "This cat is black" can be an atomic proposition if we agree that this cat and black are unanalysable simples. But if 'through the back door' we introduce mentally information such as "Black is one of 10 possible colours", then we are drawing an information-pattern comprising far more than the single element needed to represent the compresence of two simples asserted by an atomic proposition. To quote from Russell's Introduction to Tractatus, (p. 12), "It is not contended by Wittgenstein that we can actually isolate the simple or have empirical knowledge of it. It is a logical necessity demanded by theory, like an electron." Thus to demand an example of an atomic proposition in terms of experienced things may be like demanding to be shown an example of an electron. Scepticism of the 'reality' of either is logically possible; but I have no wish to objectivise atomic propositions to any greater degree than electrons.

The point is that the scientific discipline treats as elementary certain concepts, such as the mark-on-a-scale. It abstracts, from the host of (scientifically) irrelevant facts about its logical counters, statements which it treats as unanalysable and therefore self-sufficient as constituents of a total scientific proposition. Dr. Good would probably agree that neither of his sentences could retain its simple form if it were defined with proper scientific rigour.

I agree with Professor Bartlett that my present approach could not be used to determine (e.g.) the magnitude of h; I am not sure however that the same need be true of dimensionless constants. My emphasis has been on the necessity of making structural models of any relationships we may assert as a result of experimentation. The problem of defining logical elements out of which to build our structure sets certain limits to the kind of statements we can make about a given experimental situation. In particular it introduces quantization of the states of the situation-as-described.

Dr. Gabor, I think, accepts this view and has in his symposium paper deduced its consequences in the field of electromagnetic measurement. He was himself responsible for clarifying the quantal nature of structural information, though I think he would not describe that work as a deduction from quantum theory. The introduction of the metrical aspect and the logical justification of a quantal approach in my Phil. Mag. paper can be thought of as complementary to Gabor's earlier work.

Dr. Reeves' suggestion that the information-space should be mapped on a line is essentially what one adopts in computing the selective information-content or 'amount of detail' (Ref. 4) in a result, which certainly 'combines' the metrical and structural measures. Its disadvantage is precisely that it does remove the vectorial aspect and with it the concepts of relevance or bearing, proximity of information-points, and the like, which the vector-space was invented to represent.

I am not sure how far confusion is divided between Dr. Moran and myself. I agree with him on the importance of distinguishing between scientific and philosophical theory, and I wish I could feel sure that he has correctly distinguished the two components in my papers.

My chief philosophical point is a linguistic one: if the ultimate appeal of every statement in physical science is to measurement, then in so far as a statement is logically verifiable it must be quantal. Because the structure of common language is not isomorphic with the structure of "scientific" experience, we may not always make statements in the logical form to which our epistemological principles appear to commit us. The fact that we may decide the magnitude of a quantum by a convention does not however make the use of a quantal language conventional, or at least any more conventional than the use of (e.g.) subject-predicate language-forms.

To remove the sails from Dr. Moran's favourite windmill, I would insist, I think with him, that the moral is the need for a less restricted discipline if experience is to be seen "whole".

DISCUSSION ON PROF. M.S. BARTLETT'S PAPER "THE STATISTICAL APPROACH TO THE ANALYSIS OF TIME-SERIES"

DR. I.J. GOOD.

I should like to comment on two related questions among the large number treated by Prof. Bartlett, with reference to § I.1 and I.4 of his paper.

Let $\mathcal{H}$ be a hypothesis and E an event (the result of an experiment). Let $p = P(E|\mathcal{H})$, $p_0 = P(E|\bar{\mathcal{H}})$ where $\mathcal{H}$ and $\bar{\mathcal{H}}$ are rival hypotheses. Prof. Bartlett asserts, in § I.2., that it is intuitively obvious that p/p_0 (or equivalently $\log p - \log p_0$) is the best statistic to use. I agree that it is intuitively reasonable but it does not seem quite obvious to me. Why is it better, apart from its simplicity, than say $(\log p)^3 - (\log p_0)^3$? On the other hand if we are allowed to talk about the probability of a hypothesis we have

$$\text{factor in favour of } \mathcal{H} \text{ given } E = \text{dfn. } \frac{O(\mathcal{H}|E)}{O(\mathcal{H})} = \frac{p}{p_0}$$

This makes it completely convincing that p/p_0 exhausts all the information from the experiment. It seems a pity to inhibit one's freedom by refusing to talk about the probability of a hypothesis. The above proof that p/p_0 exhausts the information applies even if we regard the initial odds of $\mathcal{H}$ as unknown, and shows that Bayes' theorem is useful even if the initial odds are completely unknown ≠ - were that possible.

The rest of what I have to say may be regarded as a continuation of my remarks following Mr. Cherry's paper. Taking expectations of the equation written down yesterday we get

expected wt. of evidence (from one experiment) in favour of $\mathcal{H}$ if $\mathcal{H}$ is true

$$= -\sum_v P(E_v|\mathcal{H}) \log P(E_v|\mathcal{H}) + \sum_v P(E_v|\mathcal{H}) \log P(E_v|\bar{\mathcal{H}})$$
$$= J(\mathcal{H},\bar{\mathcal{H}}) - J(\mathcal{H},\mathcal{H}),$$

with a self-explanatory notation analogous to that of Prof. Bartlett. The second term $J(\mathcal{H},\mathcal{H})$ is the selective entropy of $\mathcal{H}$ and the first term may be called the 'cross-entropy' of $\mathcal{H}$ and $\bar{\mathcal{H}}$, though it is not symmetrical in $\mathcal{H}$ and $\bar{\mathcal{H}}$.

There is one hypothesis, $\mathcal{H}_*$, of particular interest, namely the one for which $P(E_r|\mathcal{H}_*)$ is independent of r. This may be called the 'flat hypothesis' since the graph of the chances $P(E_r|\mathcal{H}_*)$ against r is a horizontal line. $\mathcal{H}_*$ is the hypothesis for which the entropy is maximised. In statistical mechanics it may correspond to the division of phase space into cells of equal volume. It has the further property that $J(\mathcal{H},\mathcal{H}_*)$ is independent of $\mathcal{H}$, being equal to $J(\mathcal{H}_*,\mathcal{H}_*) = \log N$, where N is the number of possible values of r. It follows that if we are taking it for granted that either $\mathcal{H}$ or $\mathcal{H}_*$ is true, then

expected weight of evidence in favour of $\mathcal{H}$ if $\mathcal{H}$ is true

$$= J(\mathcal{H}_*,\mathcal{H}_*) - J(\mathcal{H},\mathcal{H})$$

= (maximum entropy) times (redundancy of information source).

For example, a language of high redundancy looked less like absolute nonsense than one of low redundancy.

---

i.e. not judged to lie in any interval shorter than $(0, \infty)$.

We see therefore that the entropy concept is strictly available for statistical inference when one of the hypotheses is the 'flat hypothesis'. (Cf. the footnote to § I.4. of Prof. Bartlett's notes).

I should not be surprised if 'cross-entropy' had an application in statistical mechanics.

DR. P.A. MORAN

Mathematical statisticians must regard Professor Bartlett's paper as being of outstanding importance for three reasons. Firstly, it draws attention to his new approach to periodogram analysis (which he has already described in a recent paper in Biometrika). In the last few years the method of analysing time series by periodogram analysis has been largely discredited partly because many time series have continuous rather than discrete spectra, partly because of the large sampling fluctuations and partly because of the absence of tests of significance. If however one is aiming at estimating a continuous spectral distribution, as for example in the analysis of ocean waves, Bartlett's smoothed periodogram provides a satisfactory method and this constitutes a most important advance. The second important contribution of Bartlett's paper is the sorting out of the relationship between the concept of information as entropy and Fisher's definition of information. This will undoubtedly have much influence on future work. Finally Bartlett draws attention to the important problem of the properties and validity of the method of maximum liklihood when used for the estimation of parameters in stochastic processes, or more generally, in dependant probability systems of any kind. A little work has already been done on this by Wald and Grenander, but it is clear that here is a large unexplored field of research with important applications.

PROFESSOR M.S. BARTLETT (IN REPLY):

In reply to Dr. Good's query about p/p_o I suggest that its relevance is "obvious, after an appraisal of the entire situation". Its appropriateness may be seen in more than one way. I certainly do not object to Dr. Good's argument, which I have used myself, although it must be recognised that one has, in consequence, moved from a statistical to a logical field of discourse. It has, however, been shown (without such a migration) that a knowledge of P/p_o exhausts the statistical evidence about the two hypotheses, any further probability conditional on the observed value of p/p_o being indifferent to which hypothesis is true.

I should like to thank Dr. Moran for his kind remarks, but on the maximum likelihood problem think it advisable to remind readers that Dr. Moran's own comment shows that I am neither the first, nor shall I be the last, to discuss this problem.

DISCUSSION ON MR.P. M. WOODWARD'S PAPER ON "THEORY OF RADAR INFORMATION"

PROFESSOR M. S. BARTLETT

I should like to draw attention to the distinction made in my paper between problems of specification and of inference. While I agree with Mr. Woodward that if the prior distribution of an unknown quantity to be estimated is known, so that the specification is complete, it is best to use his approach, in many cases this is the case. The methods of inference I have sumarized are still, however, immediately available. As examples, consider first the problem of signal detection, which is a case of comparing one hypothesis, of noise only, against the alternative of noise plus signal, so that the likelihood ratio is the relevant criterion (as pointed out in the book "Threshold Signals", chapter 7). Secondly, take Mr. Woodward's own estimation problem. Here

$$\log p = C - \frac{1}{N_o} \int [Y(t) - u(t-\tau)]^2 \, dt.$$

Differentiating one with respect to τ and equating to zero gives the maximum likelihood equation. Differentiating again and averaging gives for Fisher's information functions the exact result

$$I(\tau) = \frac{2}{N_o} \int [u'(t-\tau)]^2 \, dt = \frac{2E\beta^2}{N_o}$$

the reciprocal of which gives a lower bound for the variance of any (unbiassed) estimate.

Mr. Woodward's further detailed analysis of the posterior probability distribution is of course of considerable interest and value. I have not examined the corresponding analysis by this alternative method, but imagine that a study of the sampling distribution of the maximum likelihood estimate would yield comparable conclusions.

Mr. F. ROBERTS

It would be interesting if Mr. Woodward ever carried the analysis of Radar Information to the point where the shape of the returned pulse is analysed to give the material characteristics of the target as well as information on its movement. Experienced radar operators during the last war were able to get a surprising amount of information of this kind by listening to the echoes on a pair of headphones.

MR. D. M. MACKAY

It is interesting and may be suggestive to rewrite Mr. Woodward's expression in a form $\sigma \geqslant \frac{1}{\beta}\sqrt{\frac{N_o}{2E}}$ expressing the numbers of proper-scale intervals involved.

Since β is the transmitter bandwidth, 1/2 defines a "characteristic time interval" T for the system. The corresponding proper-scale-unit of interval ΔT is proportional to σ. At the same time $2E/N_o$ (as mentioned in the glossary) is proportional to the metron-content of the received signal, so that $\sqrt{2E/N_o}$ is proportional to the number, n_s say, of proper-scale intervals occupied by the signal amplitude.

Making these substitutions, we may write the expression for σ as $T/\Delta T \leqslant$ const. x n_s. It then appears to state simply that with a suitable proportionality-constant the number of scale-units occupied on the proper-scale of T cannot exceed the number occupied on the proper-scale of the signal amplitude. As this comes close to being axiomatic, on the basis of Bartlett's suggestion that such questions of inference can usefully be discussed from a different standpoint. In the present case it is not obvious that the constant of proportionality could have been predicted rigorously on this simpler basis, but perhaps the quickest way to advance

towards rigour in the theory of scientific information, will be to study inductively as many as possible of the similar instances which abound in physics. If conventional treatments are first used to derive exact relationships, their reformulation in the abstract terms of information theory can often indicate immediately the general principles concerned. In the light of these, such relations as the equation defining σ^2 above take on, I think, a new and simple meaning.

MR. O. E. BAILEY

RADAR INFORMATION IN TERMS OF SIGNAL VISIBILITY.

(1) Woodward has discussed the amount of information conveyed by radar signals in terms of the accuracy of range measurement. On a radar equipment with a PPI type of display the emphasis is usually on detection of signals, exact measurement of their position being unimportant. But of course detection implies location to some degree of accuracy.

At RRDE we have measured the visibility of signals on a PPI for various values of the radar parameters in terms of the signal-noise power ratio before detection. From a large number of readings with a given set of parameters we plot probability of detection against signal input power and take the level for 50% probability of detection as the minimum visable signal.

It is perhaps of interest to consider what quantity of information, in Shannon's sense, this type of measurement associates with a radar signal. The calculations are simplified if, instead of using the experimental results directly, we find a model to represent the process of detection which gives an explicit expression for visibility approximating to the experimental figures.

(2) Consider a normal pulse radar with a PPI display having the following parameters:

S^2 is the signal-noise power ratio during the pulse before detection

n is the RMS noise voltage before detection

B is the bandwith

T is the pulse length

R is the number of pulses transmitted while scanning one beamwidth

$n\bar{x}(s)$ is the mean output voltage after detection

$n^2\bar{x}^2(S)$ is the mean square deviation of the output voltage after detection

We assume the PPI area divided into a number M of cells each equal in area to the product of the beamwidth and pulse length a nd that we know that there is a signal occupying exactly one, and only one, of these cells. As a criterion of visibility we shall assume that the eye selects as the signal the cell containing the largest mean voltage averaged over the cell area. This is equivalent to assuming that the PPI is linear. We shall calculate the probability that the cell selected in this way is in fact the signal cell as a function of S.

Since the bandwidth is finite we can replace the continuously varying output voltage by a set of readings taken at intervals 1/B, and there will thus be BRT readings per cell. Then, averaging over a single cell, the mean voltage is $n\bar{x}(S)$, the variance $\frac{n^2\bar{x}^2(S)}{BRT}$ and the distribution approximately Gaussian. We then find that the probability that the signal cell has the largest mean voltage is

$$P(S) = \int_{\sqrt{\frac{RBT}{n^2\bar{x}^2(S)}}\left\{\bar{x}(0) - \bar{x}(S) + \sqrt{\frac{2\bar{x}^2(0)}{BRT}}\, f(M)\right\}}^{\infty} (\pi)^{-\frac{1}{2}} \exp(-z^2)\, dz$$

where f(M) is given by

$$\int_{-\infty}^{f(M)} (\pi)^{-\frac{1}{2}} \exp(-z^2)\, dz = (\tfrac{1}{2})^{1/M}$$

and is a very slowly varying function of M which may be taken as having the constant value 2.5 for the range of M in which we are interested.

It follows that P(S) is the probability that if we select the cell with the largest mean voltage we have found the true signal position. If we call S_o the value of S for which there is 50% probability of correct detection, then S_o is the root of

$$\bar{x}(S) = \bar{x}(0) + (2\bar{x}^2(0)/RBT)^{\frac{1}{2}} f(M)$$

and can be compared with experimental results. Substituting the known parameters of noise after linear detection and making some approximations, we find

$$S_o = 1.8 \sqrt{(RBT)^{-1} + (RBT)^{-\frac{1}{2}}}$$

Over the range of parameters for which measurements have been made this formula agrees with the experimental results to within 2 dbs, so we may use it to calculate the approximate amount of information conveyed by the signal. The simplifications involved in its derivation are of course drastic.

(3) We assume that a priori the signal is equally likely to appear in any one of the cells. The amount of information conveyed by a given signal is therefore

$$H = \log M + P \log P + (1-P) \log \left[(1-P)/(M-1)\right]$$

$$= P \log M + P \log P + (1-P) \log (1-P).$$

The maximum of this is log M and the minimum 0 as might be expected.

If we calculate the amount of information conveyed by a signal of total energy E we find that it is not independent of the way in which this energy is distributed. It depends on the values of R and T individually. In order to illustrate the way in which H varies we shall take particular values for R and T and plot contours of constant H against log B and E/N where N is the noise power per unit bandwidth and E depends only on the signal strength.

These contours for a certain set of conditions are plotted in Figure 1. The maximum possible amount of information is 0.43 E/N. This is most nearly approached when $BT \sim 1$, but the discrepancy is still of the order of 15 db. We cannot continue the curves for $BT < 1$ since the signal amplitude in the output for a given input then decreases.

(4) We have still to consider how the extra information which we obtain when we can locate the signal more accurately than to a particular cell is to be included.

To simplify the argument let us now consider an A scope display. We know that the probability of detecting a signal on an A scope is approximately equal to that on a PPI, so we can apply the results of the previous arguments, qualitatively at least.

Suppose that we have an A scope displaying a total range A. Let P be the probability of detecting the signal correctly and $Q(r-r_o)\,dr$ be the probability when we have detected the signal that a range between r and r + dr will be measured if the true range is r_o. If we make a large number of trials with a fixed signal strength and true range, we shall find that a proportion P of the measured ranges are clustered around r_o with a distribution Q and the remainder are spread out at random in range. In fact, the probability of measuring a range between r and r + dr in a given trial will be $(PQ + (1-P)/A)dr$. This will have the shape sketched in Figure 2. The priori probability is dr/A so the amount of information is

$$\log A + \int_0^A \left\{P\,Q + (1-P)/A\right\} \log \left\{P\,Q + (1-P)/A\right\} dr$$

$$+ \doteqdot P \log A + P \log P + (1-P) \log (1-P) + P \int_0^A Q \log Q \; dr$$

The first three terms behave approximately as the curves of Figure 1 and the last term is a comparatively small correction depending on the precision of measurement when the signal has been detected. If we assume that Q is approximately Gaussian with variance σ^2, the amount of information becomes

$$H \doteqdot P \log P + P \log P + (1-P) \log (1-P) - P \log \sigma (2\pi e)^{\frac{1}{2}}$$

σ will decrease as E and B increase, so that total amount of information will be represented by contours of the form sketched in Figures 3.

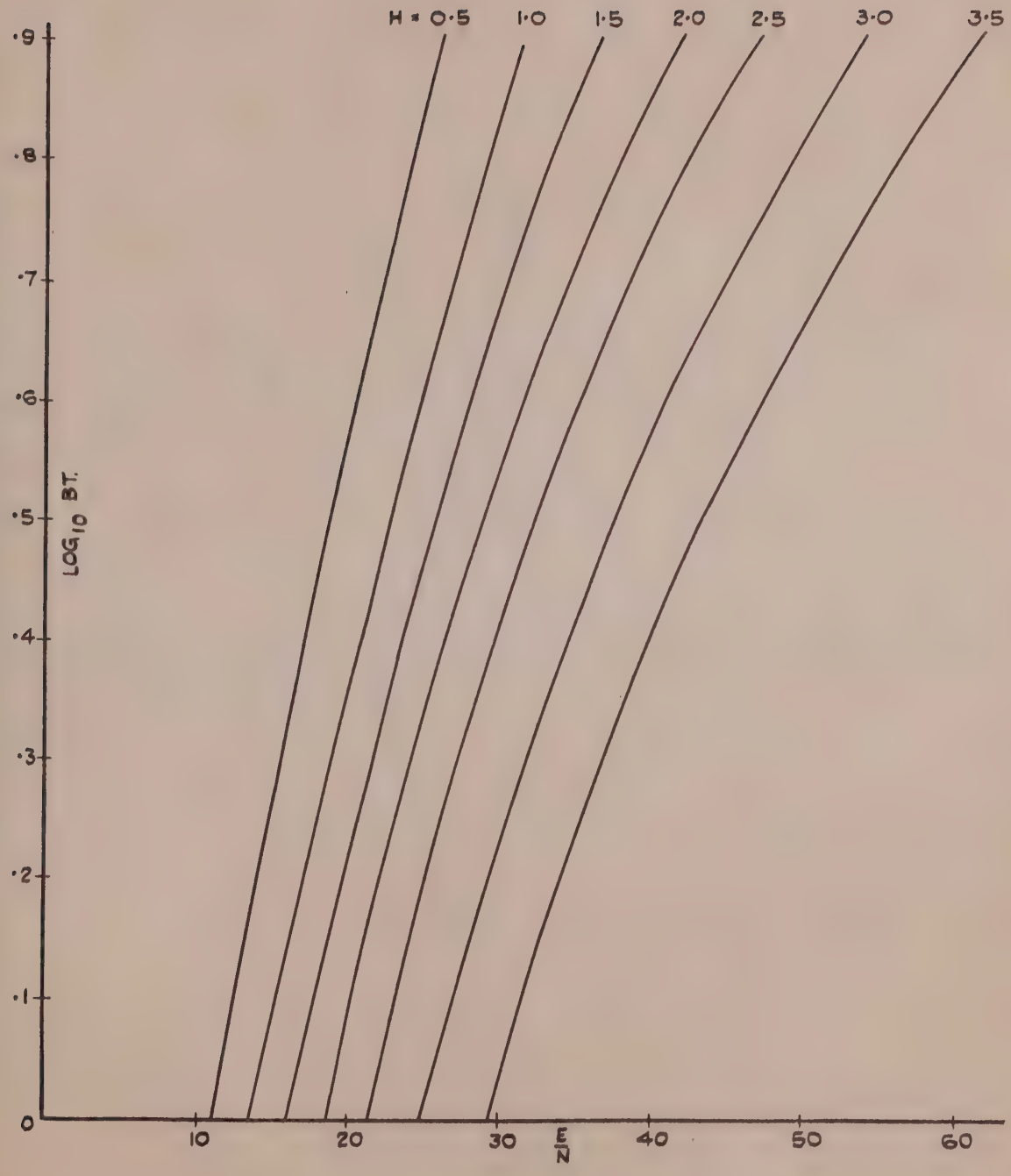
It will be seen that this represents a type of begaviour which is at least plausible. For a constant bandwidth as E/N increases the amount of information increases rapidly in the vicinity of the threshold of detection and then slowly as the precision of measurement is improved. For constant E/N increasing the bandwidth at first increases the amount of the information decreases rapidly to zero.

(Addendum)

(5) The difference between these results and Woodward's seems to the the result of measuring information at different points in the chain. Woodward measures the amount of information contained in the received waveform, without reference to any particular method of measuring range. We have specified a method and have calculated the amount of information derived from the measurement. When the signal energy is large there is negligible ambiguity and all the information contained in the waveform can be recovered in the measurement. In the ambiguous region any single measurement is likely to be wrong and the ambiguity is transformed into a negative amount of information which reduces the total derived information from the measurement to practically zero.

An important practical problem is to discover what system of measurement (i.e. display) enables the maximum amount of useful information to be derived, particularly in the ambiguous region.

FIG. 1.



CONTOURS OF INFORMATION IN DECIMAL UNITS.
R T AND M CONSTANT.

FIG. 2.

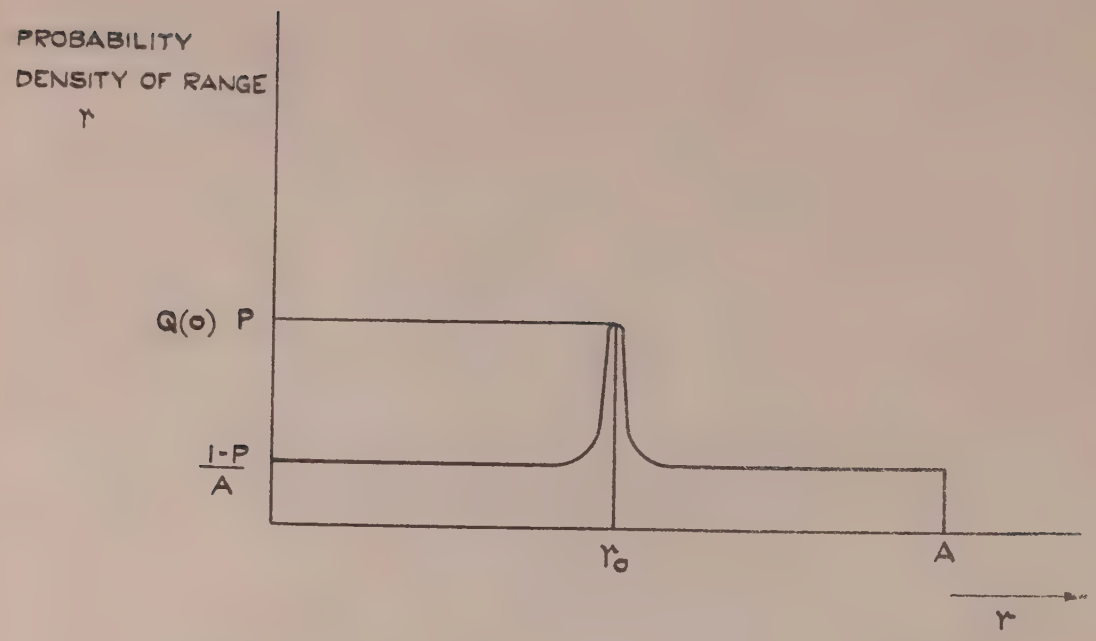
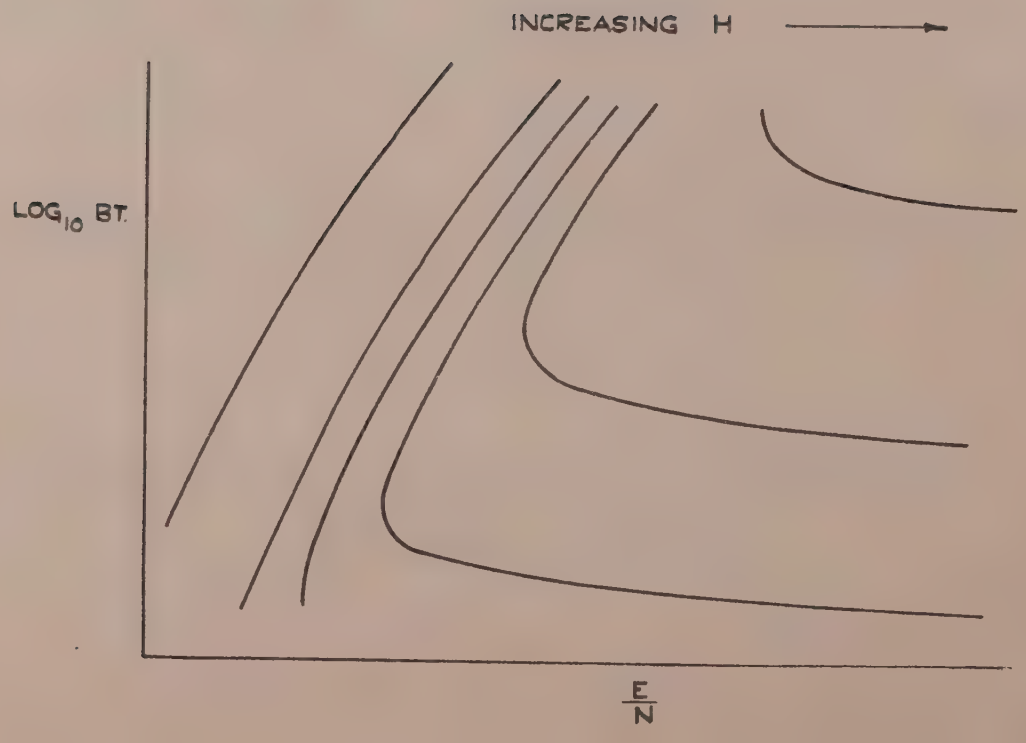

PROBABILITY
DENSITY OF RANGE
r
Q(o) P
1-P
A
r0
A
r

FIG. 3.


INCREASING H
LOG10 BT.
E
N

MR. P. M. WOODWARD (in reply):

I am very grateful for Professor Bartlett's suggestions which I look forward to examining in detail. It must be admitted that my treatment is uncontroversial provided an a priori distribution can be specified, as it always is in Dr. Shannon's theory. But as Professor Bartlett foresees, this might be impossible in some problems. If it cannot be specified, I am doubtful whether Dr. Shannon's precise concept of information has any meaning.

"Not yet" is the reply to Mr. Roberts, and I must thank Mr. MacKay for his stimulating remarks.

I hope Mr. Bailey will forgive me for not having grasped straight away the very interesting point he has raised. I used Shannon's

$$H(\chi) - H_y(\chi),$$

while Mr. Bailey has used

$$H(y) - H_\chi(y).$$

The interesting fact is that we obtain results which are obviously different, especially in the ambiguous region, yet the two formulae above are mathematically equivalent. To equate them is, in fact, to write down Bayes' Theorem in disguise. The mathematical reason is exactly as Mr. Bailey has suggested: it lies in the interpretation of y. My y is the complete received waveform, his y is an observer's best guess. My $H_y(\chi)$ is the entropy of the a posteriori distribution on any one occasion, his $H_\chi(y)$ is the entropy of best guesses over an ensemble of different occasions. The practical conclusion is this. The moment an observer throws away the a posteriori distribution and makes an exact guess instead, he destroys information. In a sense, he adds bogus information because he disguises his original uncertainty, but his sins find him out in the ensemble! I am discussing this question, and that of optimum display, in a forthcoming paper.*

*"Theory of Observation", T.R.E. Journal (January, 1951.)

DISCUSSION ON DR. D.K.C. MACDONALD'S PAPER, "FLUCTUATIONS AND THE THEORY OF NOISE"

MR. R.E. BURGESS

Equation (12a) shows that the autocorrelation function of the thermal noise voltage appearing across the terminals of a parallel R.C. circuit has the same form as the response of the circuit to unit current impulse i.e. a function decaying exporentially with time constant R.C.

This is a special case of a general result which can be established readily for any two-terminal linear passive network as follows:

Let $Z(j\omega) = R(\omega) + jX(\omega)$ be the impedance of the network at angular frequency ω. If T is the absolute temperature of the network, the spectrum of the thermal noise voltage appearing at its terminals is given by $\frac{2kTR(\omega)}{\pi}$. The autocorrelation function of the thermal noise voltage v is therefore given by

$$\psi(\tau) \equiv \overline{v(t)v(t+\tau)} = \frac{2kT}{\pi}\int_0^\infty R(\omega)\cos\omega\tau\, d\omega$$

Now the voltage response to the application of unit current impulse to the network at $t = 0$ is given by

$$G(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} Z(j\omega)\, e^{j\omega t}\, d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \left[R(\omega)\cos\omega t - X(\omega)\sin\omega t\right] d\omega$$

If now t is a positive quantity $G(-t)$ must be identically zero since there can be no response before the application of the impulse.

Hence $0 = G(-t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \left[R(\omega)\cos\omega t + X(\omega)\sin\omega t\right] d\omega \quad t > 0$

Thus $\int_0^\infty R(\omega)\cos\omega t\, d\omega = -\int_0^\infty X(\omega)\sin\omega t\, d\omega$

and $G(t) = \frac{2}{\pi}\int_0^\infty R(\omega)\cos\omega t\, d\omega. \qquad t > 0$

Hence we obtain the simple results

$$\psi(\tau) = kT\, G(|\tau|)$$

which is valid for any value of τ, and

$$\psi(0) = \overline{v^2} = kT\, G(0)$$

If the same two - terminal network forms the anode load of a saturated diode carrying a current I, the noise voltage appearing across the network due to the shot effect has the spectrum $\frac{eI}{\pi}|Z(j\omega)|^2$ and the auto-correlation function is hence

$$\phi(\tau) = \frac{eI}{\pi}\int_0^\infty |Z(j\omega)|^2 \cos\omega\tau\, d\omega$$

$$= \frac{eI}{2\pi}\int_{-\infty}^{+\infty} |Z(j\omega)|^2\, e^{j\omega\tau}\, d\omega$$

$$= eI\int_{-\infty}^{\infty} G(t)\, G(t+\tau)\, dt$$

by the convolution theorem. This result also readily follows from an extension of Campbell's Theorem when it is recalled that each electron transit produces a response of the form $e.G(t)$.

The autocorrelation coefficient $\phi(\tau)/\phi(0)$ for the shot noise voltage is not therefore in general the same as that for the thermal noise, $\psi(\tau)/\psi(0)$. One special case in which they are identical is when $G(t) \propto e^{-\alpha t}$ since then $\frac{\phi(\tau)}{\phi(0)} = \frac{\psi(\tau)}{\psi(0)} = e^{-\alpha\tau}$ This is the case of the simple RC network. The general class of network for which this is valid is that in which the shunt conductance $g(\omega)$ is independent of frequency, for then

$$|Z|^2 = g\,R(\omega)$$

whence $$\phi(\tau) = \frac{eIg}{2kT}\psi(\tau)$$

This shows that the mean square shot noise voltage is $\frac{eIg}{2kT}$ times the mean square thermal noise voltage and also that the autocorrelation coefficients (and hence spectral shapes) are identical.

PROF. P. GRIVET
DR. J.L. STEINBERG

We wish to draw the attention of the Symposium to a new source of noise which seems to be important in very large bandwidth linear amplifiers. It concerns "gain fluctuation noise" already mentioned by Dicke (Rev. of Scient. Instr. 17, 1946, p.268), the study of which has been undertaken in our Laboratory (J. Nosnier and J.L. Steinberg, C.R. Acad. of Sciences, 230, p.438-440).

We think that an unknown mechanism brings the high level of "flicker" noise into the signal amplified by a high frequency linear amplifier. The importance of this noise seems to be particularly striking in large frequency bandwidth receivers (Bandwidths larger than 5 Mc/s).

In the case of the output power spectrum of a 2 megacycles bandwidth receiver operating on 30 megacycles, the abnormal noise component on 10 cycles per sec. may be twice as large as the normal noise component on the same frequency. This order of magnitude cannot be explained by the simple theory assuming a variation in the transconductance of the tube produced by the "flicker effect".

DR. I.J. GOOD

What exactly is the relationship between white noise and ordinary white light from the sun?

DR. D.K.C. MACDONALD (in reply):

Mr. Burgess' discussion of the general equivalence of the differential equation for $\psi(\tau)$ to the current-impulse response of a passive network is very welcome. I had considered it worth while here to deal with the detailed stockastic analysis of an elementary circuit in order to bring out clearly the nature of the specific physical assumptions involved.

In his interesting consideration of the relationship to the shot-noise correlation function one might perhaps point out that it is there tacitly assumed that the "transit-frequency" of the diode ($1/T$ say) is large compared with any significant contribution of $|Z(j\omega)|^2$ to the integral for $\phi(\tau)$. The other obvious special case of the equivalence referred to is that of an idealised tuned circuit with only "parallel damping". The close of this analysis should also read:-

".... $\left|Z\right|^2 = \frac{1}{g} \cdot R(\omega)$ (where g is a constant)

whence $\phi(\tau) = \left(\frac{eI}{2kTg}\right)\gamma(\tau)$

This shows that the mean square shot noise voltage is $\frac{eI}{2kTg}$ times ...."

In reply to Dr. Good, "white" noise in, say, a radio amplifier is that which has a uniform power spectrum over the relevant bandwidth.

On the other hand, the radiation emitted from the sun has the maximum of its power spectrum (according to the Wien displacement law) in the visible spectrum ($\lambda \sim 4,800A$) and therefore the power spectrum is more or less flat over the range of visible light.

DISCUSSION ON DR. T. GOLD'S PAPER on "HEARING"

PROF. B. VAN DER POL.

In connection with the propagation of pulses along the aural nerve, I asked the physiologists the truth of the following observation (and they confirmed it). When a low audible note is impressed acoustically from the outside on the ear, the frequency of the nerve impulses may correspond with the frequency of the acoustical note, which may be either of a sinusoidal form or consisting of equally spaced impulses. When thereupon the frequency of the acoustical note is gradually raised, the frequency of the nerve pulses goes up as well till the moment when a frequency, critical for the nerve, is reached, at which moment the frequency of the nerve pulses falls to exactly half the frequency of the external excitation. When the frequency of the external excitation is still further increased, the nerve pulses follow the external frequency an octave lower till again the critical frequency of the nerve pulses is reached at which moment the frequency of the nerve pulses falls to 1/3 of the external frequency.

Here we have before us a beautiful example of frequency "demultiplication" which is the typical phenomena associated with relaxation oscillations. A similar frequency demultiplication occurs occasionally in the beating of the heart where it is called "heart block". Surely these phenomena are of an essentially non-linear character, and I often wonder whether the normal functioning of the ear could well be described by linear differential equations.

DR. T. GOLD (in reply):

The frequency demultiplication which is observed in the auditory nerve is in no way unique. It is the inevitable consequence of the stimulation of a nerve fibre, possessing the normal type of time-constant of recovery, with an oscillatory signal. One is hence not forced to attach any particular significance to this effect in the ear.

In the higher part of the auditory spectrum, where the fluctuations of sound are much faster than the nerve impulses, this effect is known to break down, as one would expect. Furthermore there is much direct evidence that frequency analysis has already taken place before the conveyance of the information through the nerve.

This action would hence be insufficient for the basis of a theory of hearing, and it is also not necessary, in the light of present knowledge, to invoke any such special method of transmission in the auditory nerve

DISCUSSION ON DR. W. E. HICK'S PAPER "INFORMATION THEORY IN PSYCHOLOGY"

NOTE BY DR. HICK

Professor Adrian referred to me a question concerning the total storage capacity of the human brain. As far as I am aware, no serious attempt has been made to estimate this quantity, in the general sense. It is, of course, possible to estimate how much information in a specific form is held at a given time - e.g., a child of a certain age is expected to have a certain vocabulary. But there is also information contained in social behaviour, skills, etc., which would be very difficult to measure, though not impossible in principle. Moreover, a large part of the information that can be elicited from a person is, in effect, a transformation of information supplied. In this case, clearly it is in the form of the transformation that the stored information must be sought. Such a transformation resembles a sub-routine, which may be used for many different purposes. The "Unconscious" of Psychoanalysis amounts roughly to a system of sub-routines, and even in a healthy person, it may take months to clarify their true nature. The apparently more straightforward approach by enumerating all the distinguishable actions a person can perform repeatably on demand runs into similar difficulties. In short, we can find a lower limit, but I doubt whether any psychological procedure could indicate the upper limit, unless the question were far more exactly and restrictively defined. The physiological approach - e.g., the possible combinations of nerve cells - may ultimately be more successful, perhaps in the manner hinted at in the abstract of Dr. Grey Walter's paper.

DISCUSSION ON DR. J. A. V. BATE'S PAPER "SIGNIFICANCE OF INFORMATION THEORY IN NEUROPHYSIOLOGY

MR. F. ROBERTS

A possible hypothesis for learning is that it is the optimal exploitation of the Hartley Shannon law. The memory would then be merely the store and the rest of the nervous system an information transformation machine, whose efficiency could then be directly related to the Hartley-Shannon law.

DISCUSSION OF DR. A. M. UTTLEY'S PAPER "INFORMATION, MACHINES AND BRAINS".

PROFESSOR B. VAN DER POL

It is well known that in modern digital electrical computers a "memory" is built into the machine, and the designer is perfectly aware of the number of "bits" which can be stored in this artificial memory. Now Wiener and others have compared this artificial memory with the human brain. I think, therefore, that it is a natural question for a mathematician to ask the distinguished psychologists present whether they can tell me the order of magnitude of the number of "bits" which, at a given moment, can be stored in the human brain.

The tentative answers given in the meeting varied between 10^9 and 10^{20}. However the fact should be taken into account that much information is usually stored in the brain which only under hypnosis can be brought to light.

MR. A. M. TURING

Concerning learning machines, I agree that the conditional transfer could probably be made a basis for learning, but would like to point out another feature which could also be used. If, as is usual with computing machines, the operations of the machine itself could alter its instructions, there is the possibility that a learning process could by this means completely alter the programme in the machine. In reply to a question by Mr. Rey on this point - by learning based on conditional transfers the machine could only reach 2^C different states of training, where C is the number of conditional transfers originally in the machine, whereas with 'learning by altering instructions' the number is 2^S where S is the storage capacity of the machine (in binary digits). With the Manchester machine, for example, S/C could not be less than 20.

I wish to expand on a point touched by Dr. Uttley. In many types of investigation, e.g. in the theory of information, it is a useful assumption that 'computation costs nothing'. It is important however not to let this assumption become a belief. In particular when one is considering brains and computers the assumption, and the theories based on it, are not applicable.

I want also to add some remarks concerning the usefulness of a random numbers generator in a computer. The application I have in mind is that in which it is made to facilitate 'scanning' processes, i.e. searching through some set for a member with a given property. Suppose for example that one wishes to find a particular integer, less than 1000 and such that it is equal to the sum of the first thirty digits after the decimal point in the decimal expansion of its reciprocal. This problem has been chosen as one which can most quickly be solved by trial and error. There is a question however as to how the trial and error should be organised. One possibility is the systematic search, trying first 1, then 2, etc. There is also the possibility of the random search in which numbers are chosen at random and tried, and no record is kept concerning which have been tried. For the problem in question there are better methods, e.g. one knows that any solution is likely to be near 150. But let us imagine that there is no particular reason for believing a priori that one value is more likely to be right than another. Then the systematic method has the advantage of involving less testing. The extent of this advantage is measured by the equation:-

(Average number of tests on random methods) = $(\frac{S+1}{S}$. Average number on systematic method)

where S is the actual number of solutions. If S is at all large this advantage is not great. The random method has the advantage that no record needs to be kept. This can be of importance when the set being searched is of greater complexity, e.g. if it consists of triples of positive

integers. The random method is also much the easier to organise if a number of workers are assisting in the search. It is interesting to consider evolution as an example of such a random search. Breeding organisms are to be imagined as searching for improved combinations of genes. The genes are changed by a random process and the searchers work independently. Once a solution is found the searchers soon cease to search: they will have been eliminated by natural selection.

MR. H. C. CALPINE.

It seems advisable to try and make explicit an idea which has been implicit in much of this discussion, when the words "success" and "failure" have been used to describe the result of carrying out a particular instruction in the programme of a machine. Computing machines compute because they are instructed to do so by a programme, which, at least in its initial form, has been supplied from an external and human source. If we attempt to use computing machine vocabulary to describe the operation of the brain, then we are led to ask, "why does the brain compute at all?" and"How is the programme of the brain established?"

The facts which we need to describe in order to provide an answer to these questions seem to me to be, that human beings have "built-in" emotional drives which stimulate continual interaction with the outside world, and provide an initial set of values by which the results are judged.

The corresponding features of a Computer might be a random element in programme selection, together with a set of maximising principles by which the results of the operations are ordered in a scale of success and failure, and a mechanism by which successful programmes are retained for future use in similar operations.

The part of the computer programme established by this means would correspond to learning by the experimental approach to the outside world. Other parts of the programme of the brain are presumably contributed by "built-in" features, and by matching with other brains through the processes of education.

MR. D. M. MACKAY:

I wish to make two points connected chiefly with Dr. Uttley's paper.

Firstly, in reply to his criticism of Fisher's statement that inductive inference is the only method by which we acquire new knowledge, it should be said that the examples given by Dr. Uttley were of deductive and not of inductive inference. It is true that a digital computer (when working normally) can only produce conclusions which are logically tautologous, since it performs only deduction. But in inductive human inference, a hypothesis not tautologously implicit in the data is formulated by a non-deductive process, and if by successive tests it is found to acquire high probability - i.e. if the contrary null-hypothesis is not contradicted - it seems reasonable to say that "new information" has been acquired. This is however essentially of the nature of structural information. It does not indeed add to the total of observed events, but has provided a new way of classifying and showing relations between them. There is thus no contradiction with Fisher's own axiom, that no amount of manipulation could increase the "amount of information" in his (Metrical) sense, contained in a body of data. The difficulty once again is only terminological.

Secondly, in connection with the differences between machines and brains, I venture to repeat a suggestion which seems increasingly to meet suggestive experimental support. It appears likely that most of the deficiencies of reasoning-mechanisms as compared with human brains could in principle be removed, if human thinking were treated as a stocnastic and not a deterministic process. In place of switch-controlled dichotomies, one would then substitute choice-mechanisms for which only transition-probabilities could be specified.

Such an approach seems to correspond more closely with experience. In practice it is seldom that unsuccessful courses are totally eliminated, as they were for example, in the case of Dr. Uttley's maze. Normally one judges courses only to have greater or less probability of success, and so forth. 'Learning' in such a mechanism would consist in the continuous modification of transition - probablities as a result of experience, so as to make less-frequently successful courses less likely to be chosen in future. Prejudice, preference, and other logically disreputable but typically human characteristics find ready analogues, and spontaneous "mental activity" of various non-trivial kinds could be produced by an instrument designed intelligently on deliberately non-deterministic lines. Such a device would lack the philosophical barrenness which presumably attracts many to deterministic mechanisms. It would indeed raise questions of considerable philosophical interest, though in my view none which are not already implicit in the profound concept of randomness.

It is of interest to consider the application of such ideas to the design of a chess-playing machine. By automatically modifying its transition-probabilities retrospectively according to successes or failures, such a machine could in principle learn to play successfully even without prior instruction as to the best moves, - as perhaps many of us learned to play. If two such machines devoid of initial instruction except as to rules, were set to play each other at electronic speeds, it is amusing to speculate on the degree of competence to which they might eventually attain.

DR. J. F. SCHOUTEN.

Signal transmission theory has greatly benefited from the concept and quantitative formulation of "information". It enables us to tell what and how much a message contains. It further provides a yard stick for the design of optimal transmission systems.

A problem of a very similar nature arises in the design of switching, computing, coding machines etc. Here we wish to realise the desired performance with a minimum amount of switching elements. This necessitates the definition of what such thinking machines do and how much they do.

Essentially, a thinking machine receives a message (information) at one end, operates upon this information and then delivers another message at the other end. Thus the dialling by a telephone subscriber ultimately leading to the ringing of the desired subscriber, the setting of a mathematical problem to the printing of its solution, and the input of a message in a certain code to the output of a message in a different code.

Since the word "operation" is rather overloaded with meaning we suggest the word "manupilation" to define the behaviour of the thinking machine.

This leads to the following definitions:

(1) A thinking machine manipulates information.

(2) The manipulation of a thinking machine is the base two logarithm of the possible configurations of that machine. This may also be called the "order" of that thinking machine.

(3) The manipulation of two thinking machines working in conjunction is the sum of the manipulations of each machine.

(4) The mathematical unit of manipulation is one binary choice. The physical unit of manipulation is one binary switch.

These definitions, similar to those in information theory, provide a quantity which is additive and the unit of which as in information theory, has a simple mathematical and physical meaning.

The manipulation of multiple stop switches should be defined in terms of the number of binary switches performing the same manipulation.

Further studies lead to similar aspects of redundancy as in information theory.

The purpose of a memory incorporated in a thinking machine may often be interpreted as a means to reduce the total manipulation needed for the given task, including the added manipulation needed to consult this memory.

A problem arises in view of the fact that a switching machine is not only determined by its number of independent switches, but also by the number of contacts per relay. The number of contacts evidently does not affect the possible configurations. It is, however, a well known practice to reduce this number by introducing further relays. This indicates the danger of identifying in a given machine the total number of binary switches with the manipulation of the machine. The manipulation is at maximum equal to this number but often considerably less. The manipulation or order of a mouse-trap is one, that of the human brain not bigger than ten milliard.

DR. R. A. FAIRTHORNE

I should like to mention some consequences of applying information theory to computing. So far from 'digital' and 'analogue' devices being distinct types of computer, in fact they differ only in the intensity or 'temperature' of the information they process. Because second and higher derivatives of physical magnitudes involved are bounded, and there is a threshold to discrimination, only a finite discrete amount of information can be extracted from a given output, however 'continuous' its physical apperance. That is, if a certain number of measurements be made on the output, all other measurements can be deduced therefrom without reference to the input of the instrument.

Epistemo-dynamic efficiency parallels thermodynamic efficiency in that, other things being equal, information processing should take place at high intensity, i.e. in binary digits with largest possible digital capacity. Thus 'step-up' transformers are needed to bring low-intensity (e.g. graphical) data to digital form. Also 'step-down' transformers are needed at the output whenever human monitoring is required, for high intensity information such as numerical table are not readily apprehended as a synoptic whole by the human mind, which for this end prefers graphical display.

A machine with step-up for input, and step-down for output is Maddida (Magnetic Drum Differential Analyzer) of Northrop Aircraft. Some high speed digital computers have step-down transformers to give graphical display of output on a cathode ray tube.

In all but cut and dried mathematical calculations, feed-backs from observer to machine and to input are needed to modify the programme on experiment in light of observed output; e.g. to give the orders "stop", or "continue until .....". Increasing cost of experiments make such sequential techniques imperative, and they in turn demand processing and interpretation of the measurements concurrent with the experiment. Thus information transformers are essential links in the computing chain if high speed calculators are to be used for anything other than table making.

At present, automatic step-down transformers involve economic rather than engineering difficulties. Purely automatic step-up transformers, which involve complicated judgements on pictorial contexts, are beyond the immediate horizon. Human beings remain the only available step-up transformers for general use, and that is the weakest link, technically, economically, and sociologically of computing experimental projects. But even now much can be done by co-operative design of experiment, measurement, and computation before starting work, and by cutting out unnecessary one-to-one transformation of data (in other words, copying). The residue of unavoidable human labour can be such lightened by automatic punching of readings into punched cards, equipment for which is now commercial.

Even without such aids, or with simple devices such as carbon papers, much can and should be done by reasonable office management based on informational analysis of the efficiency, relative to loss of information of the 'epistemo-dynamic' cycle used.

DR. UTTLEY (in reply):

Regarding Dr. Turing's first point on "learning" by machines. This certainly can be achieved by both the conditional transfer instruction and by the ability to modify instructions. In the example of learning a maze, both methods were employed; the instruction to turn left at a junction was modified by means of a conditional instruction depending on success/failure at some other point in the programme. I would suggest that both methods will always be necessary; the powerful ability to modify instructions will always be controlled by some success/failure condition.

Dr. Turing's second point was most important, that though a coded message may contain the same information as the original, yet they are not the same. They differ to the extent to which computation or "Manipulation" occurred, and this does not cost nothing. Here Dr. Schouten's contribution is valuable, that computation or manipulation can be analysed into elements and measured, as can information. I do not like the definition of this element as a binary choice; I prefer to analyse computation into:-

(a) Storage. The element is one stored bit.

(b) Comparison. The unit is the AND comparison.

If all stored bits are available, together with their complements, then all logical comparisons can be derived from AND. A relay operating a single pair of contacts performs this comparison since the first contact is energised only if the second is energised AND the coil is energised. Similarly the cathode of a cathode follower circuit will rise only if the anode AND the grid rise.

With reference to Dr. McKay's contribution I regard Pattern Perception as a clear example of Inductive Reasoning. If a number of objects are perceived to possess a common pattern then a general law has been discovered about them. Any assumption that this law applies to other objects as yet unobserved is an act of faith not of logic. It is only too easy for both man and machine to act on such an assumption.

On Dr. McKay' second point the introduction of transition probability into the operation of a machine is most certainly a further step towards the imitation of human thought, and this point was brought out by Dr. Good on the first day of the conference. I would only point out that this added technique can be introduced in existing general purpose digital computers. I would very much like to agree with Dr. Fairthorne that in this context and indeed in many others, comparison of analogue and digital computing methods is irrelevant. They differ only in the technical methods they employ and not in any basic way as to their capabilities.

DISCUSSION ON DR. E. SLATER'S PAPER ON "STATISTICS FOR THE CHESS COMPUTOR AND THE FACTOR OF MOBILITY"

DR. D. K. C. MACDONALD

In so far as a mental analysis of the feasibility of a machine to "play" Chess may assist us to decide more clearly what true constructive thought is and whether brains in principle are different from machines, I wholeheartedly support such work. I feel however that the actual construction of such a machine in all its relatively vast complexity - apart from personal satisfaction which it may provide - is of dubious value. To me, the purpose of Shannon's paper in Phil. Mag. was the analytical demonstration that a machine could simulate a process normally regarded as essentially "brainy" - not that one should therefore make one. I am reminded that Professor Simon once remarked to me that one could presumably make a machine that would smoke tobacco - cui bono?

DR. F. L. STUMPERS

I agree with Dr. MacDonald that the construction of a chess-machine in itself is of doubtful value. I discussed Dr. Shannon's paper with Dr. Euwe, former world-champion and mathematician and we thought that it was doubtful whether even this complicated machine would play a really good game of chess (better than an average club-player). It is very difficult to see what inferences can be drawn from Dr. Slater's interesting statistical analysis. To stress the mobility too much would bring us back to the classical school in chess (Tarrasch). In the 1920's, e.g. Reti, Euwe, Alekhine have shown that mobility and space also give a great responsibility. The Indian games rose in popularity and Alekhine played his defence in which a knight moves over the board without any increase in mobility for his side. However, when many other things are taken into account, mobility is also of importance.

MR. A. M. TURING

I wish to make two points concerning Dr. Slater's paper. I was greatly interested by the statistics provided, but fear that some people might draw invalid conclusions from them. It might for instance be thought that a good way of playing is to maximise one's mobility at one's next move, or perhaps to minmise that of one's opponent at his next move but one. It is evidently not feasible to foresee mobilities many moves ahead. Although the immediate mobility is a useful measure of the relative advantage of the players in normal play it by no means follows that it is wise to direct one's play to maximising such a measure. To do so would be like taking a statistical analysis of the laundry of men in various positions and deciding, from the data collected, that an infallible method of getting ahead in life was to send a large number of shirts to the wash each week.

I wish also to put forward a plea for a more experimental approach to chess machines. A more realistic attitude to these machines can be reached by making them than by talking about them. There are three forms which a chess machine might take.

(a) Special machines built to play chess and for no other purpose.

(b) A digital computor programmed as a chess machine

(c) Paper machines

The special machines I do not recommend. They would be too much trouble and Dr. MacDonald's comparison with a 'smoking machine' is very apt in connection with them. To programme a computer would be more rewarding, but not everyone has a computer. I wish however to recommend the 'paper chess machines' to the symposium. By making such a machine I mean laying down a definite set of rules governing play and then obeying them, or in other words 'programming oneself as a chess machine'.

Anyone can do this: the only apparatus required is board and men, and perhaps pencil and paper. Both Dr. Shannon and I have many others. I recommend it as easy, instructive and entertaining. Similar experiments can be made with 'paper learning machines'.

DR. I. J. GOOD.

I agree with Dr. Turing that mobility by itself is not enough. It seems not to take sufficiently into account the essential difference between strategy and tactics. I am reminded of the alleged motto of the Hampstead Chess Club: 'Never miss a check, - it might be mate'. Mobility may be sufficient evaluation of a position for strategic purposes, but it does not allow sufficiently for tactics, i.e. combinational play or forced variations. A good chess machine would have to analyse all forceful variations, with a suitable definition of a forceful variation. This fact has almost certainly been recognised by Dr. Shannon and others. The definition of the forcefulness of a move would have to depend partly on the difference between the strengths of an attacked piece and the attacking piece.

The analysis of the forceful variations in any given position would be a stochastic branching process, and the end-points of the corresponding tree would be quiescent positions which would need strategic evaluation (e.g. by means of a measure of mobility). The total time taken to cope with such a tree would be roughly proportional to the number of individuals in this tree, and it is known that the probability distribution of this number is extremely skew. (See, for example, Proc. Cam. Phil. Soc., 45 (1949) 360-3.) The time taken to analyse a position might be anything from 0.1 seconds to 100 years. There would therefore be a danger of the machine getting into time-trouble. It would be advisable in fact for the machine to take into account the state of its chess clock. It would then modify its definition of the threshold forcefulness of a move. In other words the programme would contain a parameter, τ, which would say how forceful a move must be in order to be examined as a sub-variation i.e. it determines whether a position is quiescent; and τ would increase when the machine was short of time. Alternatively τ might be an increasing function of the number of moves 'deep' of any analysis (i.e. a function of the generation number in the tree mentioned before).

A really clever machine would also take into account the state of its opponent's clock. The machine might move faster when its opponent was in time trouble in order to give its opponent less time to think.

DR. E. SLATER (in reply):

Dr. Good's remarks seemed to me important and very much to the point. I am sorry if I gave the impression that tactical considerations were of little significance; they would have to be taken full account of in programming the computer, and some method would have to be devised for making the machine sensitive to threats so that it would analyse forcing variations to a depth of five or six moves or more. The control of the time factor is a most interesting suggestion. , Dr. Turing's argument by analogy from what a naive laundry worker might conclude about ways of becoming rich really amounts to the suggestion that strategic advantage is the cause rather than the product of an advantage in mobility. I do not think that this can be accepted. An advantage in mobility usually appears in a game a number of moves before strategic advantage is detectable in other ways; it seems to be ann essential aspect of what chess-players understand by "development"; and it supplies the decisive criterion of winning or losing. The relegation of mobility to a place of almost trifling importance by the modern school of players, following the work of Reti, to which Dr. Stumpers has referred, has, of course, been an interesting evolutionary change in the style of the game; but one does not have to regard it as final. There are such things as fashions in chess as in other fields into which an element of intuitive appreciation enters.

The final, or rather the first question, which is raised by Dr. MacDonald, namely whether this whole discussion is not so much beating of the air, is to some extent one to which there is no answer. In so far as a present answer can be given, that has already been done by Shannon in the introductory paragraphs of his paper. The real answer would appear only after such a hypothetical machine had been built. To the general implication of Professor Simon's witticism one might reply with Faraday's classic retort: "What is the good of a new-born baby?"

DISCUSSION ON J. H. WESTCOTT'S PAPER "CRITERIA OF PREDICTION AND DISCRIMINATION".

DR. I.J. GOOD.

I may have misunderstood the point, but I would have thought that the weighting function ought to be more like $1/(t + K)$, rather than t, since the immediate future is more important than the remote future. The weighting function t gives more weight to the remote future.

PROF. M.S. BARTLETT.

In view of the trend of some of the discussions, I think it advisable to emphasize that Wiener's prediction theory, for stationary time-series affected by noise or other randomness, gives a minimum prediction error which was quite independent of any limitations due to physical apparatus. ⁂

⁂ See, for example, equations (31) and 32) of my own Symposium paper "The Statistical Approach to the Analysis of Time-Series".

MR. J.F. ATHERTON

While I accept the latter parts of Mr. Westcott's paper, I suggest that his argument that it is the physical limitations of unstable amplifiers which prevent complete prediction is unsound. Taylor's theorem is only applicable to a function if it and all its derivatives are finite and continuous, and such functions are completely predictable. We are in fact only interested in functions which do not obey Taylor's theorem.

MR. D.J. MYNALL

Mr. Westcott's paper proceeds on the assumption that an unstable network is "unusable" and that restriction to stable networks is "the only practical course".

This limitation may be too severe. Suppose we consider a particular unstable network. In general, it will be possible to devise a modification to the network by switching (electronic or otherwise) which, if effected, would result in the modified network reverting to a definite, stable state, independent of the previous condition of the unmodified network. The process of switching to the stable state is sometimes called "quenching".

Now let us specify sufficiently the class of possible excitations which the network may experience in its intended application, including thermal agitations and like disturbances. Further, let us choose a definite interval of time. Then provided that the network is never released from the quenched condition for a time exceeding the pre-assigned interval, it is possible to find limits which will not be exceeded by the currents and voltages in the various branches of the network. Having determined these limits, it becomes clear that the practicability of constructing the network so that it will operate for at least the pre-assigned time is merely a matter of engineering.

Thus, an unstable network may be used in practice, provided that its action be interrupted sufficiently often by quenching periods, during which its "memory" of previous response is erased. This erasure does not necessarily mean that the result of the response during an active period is immediately lost at the end of the period, since it may be transferred in some form to a co-operating circuit. A well-known example of the use of these principles is the linear super-regenerative amplifier.

It would seem within the bounds of possibility that the use of unstable networks in the way described might have a bearing on the problems to which Mr. Westcott has addressed himself in the present paper.

MR. R.H. BARKER

One speaker has pointed out that it may not be necessary to reject out of hand filters for prediction which are themselves unstable. It is, in fact, possible to use such an unstable filter as a component of a system which has overall stability because the unstable element receives a correction from time to time to prevent its output from increasing to saturation. A case in point is the use of a sampling servomechanism to predict the next term of a discrete time series. The arrangement is as Fig. 1.

The input to the system is a sequence of values known only at the instants $t = k\tau$ where τ is the data period. Owing to the delay of τ in the feedback loop the servomechanism operates to drive the output at $t = k\tau$ to be equal to the input at $t = (k-1)\tau$. Without special precautions such an arrangement is completely unstable. It may however be stabilised by including in the control of the servomotor a filter which is itself unstable to an extent which in some sense matches the instability of the system in its absence. For example, if a system were designed for zero velocity lag, it would predict accurately for as long as the velocity is constant. The necessary stabilising filter then has the property of generating, in response to a single impulse of amplitude A at $t = o$, the following sequence starting at $t = 0$:-

$$A, -2A, 4A, -8A, \ldots \text{ etc.}$$

A possible physical realisation is a resonant circuit of natural period 2τ combined with a negative resistance.

The overall system is quite stable because the stabilising filter receives a corrective impulse at every sampling instant.

MR. H.C. CALPINE

Mr. Westcott has suggested in his paper that the keynote for this discussion should be "that the value and importance of a criterion is tempered by the difficulty of its application as well as by the quality of its achievement when applied". This is a statement which those of us who have occasion to attempt prediction on time series generated outside the controlled conditions of the laboratory will heartily agree. I should like to suggest two ways in which practical applications often require a change in the formulation of the problem. In the first place we would often be willing to accept a prediction which is a near-optimum, meaning by this term that the extra error introduced by the departure from the true optimum is small compared with the irreducible error of prediction. In the second place a modification of the specification of the problem is often necessary. The Wiener approach to the prediction problem requires that the power spectrum of the series studied should be known a priori. There are many cases in which the a priori information is much less precise than this. For example, suppose we wish to predict the roll or pitch of a ship in a rough sea. The effective spectrum of the waves which the ship is meeting will depend on the distance and bearing of storm centres over the oceans, and the speed and heading of the vessel. The net effect is that over periods of interest the series represents a nearly stationary process, but the actual spectrum may take widely differing forms between limits of say 2.5 seconds and 36 seconds period.

The nearest formal problem will be the prediction of a stationary time series, whose power spectrum is specified only as lying between given limits. A natural approach to this sort of problems is to try to combine the predicting mechanism with an analysing mechanism which, so to speak, determines the spectrum as it goes along from the input with which it is supplied. An arrangement of this sort which is perhaps the simplest which can be conceived and can fairly readily be realized is as shown in Fig. (2). For convenience in description, the time series is supposed to be represented by an electrical voltage, called the input signal.

We arrange a set of filters, which may be single tuned circuits, whose bandwidth and spacing are such that the parallel group give a transmission path between input and output which is substantially uniform over the band of frequencies which may be contained in the signal input.

Suppose now we consider a particular input signal having a spectrum lying within the specified limits, and consider what processes it will be necessary to perform in the boxes marked "predictors" in order to obtain an optimum prediction at the output, using the mean square error criterion. It can be shown (probably most simply by considering the impulse responses of the filters) that <u>if</u> the output is a faithful copy of the input when the predictor boxes are short-circuited, then the optimum prediction over any given time can be obtained at the output by making each predictor box optimum for a signal which is the input signal modified by the particular filter to which the box is connected. Reverting to the more general case in which the input spectrum is not known in detail, we can surmise that we shall still obtain a near-optimum prediction if the input spectrum varies only slowly over the band passed by each filter, and the predictor box for each filter is that appropriate to a <u>white</u> noise input modified by that filter. The practical requirement will be, that the more detail there is in the power spectra of possible inputs, the more filter channels we shall need to make our predictor near optimum.

In what follows, the multi-channel method of prediction will be called "synthetic prediction". When the input power spectrum is known, "synthetic prediction" is not required and a predictor can be built which satisfies the mean square error criterion. In the comparisons which follow this arrangement will be called "optimum prediction".

It is not difficult to make the ideas which have been suggested above precise, and to obtain a quantitative expression for the additional error introduced by employing "synthetic prediction" rather than "optimum prediction" for some possible input. However, a better idea of the performance which can be realized may be obtained from Figure 3. This compares optimum and synthetic prediction for a case in which the power spectrum of the input, though reasonably well covered by the (two) filter channels is rather far from uniform over each filter pass band. The input was derived from a broad-band noise source feeding a tuned circuit with a Q of about 6. The synthetic prediction filters were tuned circuits ($Q \simeq 5$) overlapping in response at approximately the peak of the input spectrum (Fig. 3).

Each synthetic prediction filter was followed by a predictor box appropriate to that filter when fed by white noise, and the combined output was recorded. The same input was fed to an optimum predictor whose output could also be recorded.

Figure 3 shows firstly the results of optimum prediction over time intervals of 0.2, 0.45, 0.9 periods of the tuned circuit determining in input spectrum. The second group of records in Figure 3 shows superposed the results of optimum and synthetic prediction over the same periods. It will be seen that the two predictions are in fair agreement though there are occasional gross discrepancies. The third

record shows the error introduced by the filters alone, i.e. it compares the input with the "synthetic prediction" for zero prediction interval. By switching the signal on and off the transient errors introduced by the filters in the synthetic predictor are also recorded.

In conclusion it may be pointed out that the condition in which observation is "switched on" at a given moment, and prediction is required to be continuously optimum (or near-optimum) thereafter, is also a case of frequent practical importance and also requires an extension of the current theory.

MR. J.H. WESTCOTT IN REPLY.

Dr. I.J. Good's contribution is interesting as it raises the fundamental issue of how one judges prediction. He suggests that a weighting function $1/(t+K)$ would be most suitable since the immediate future is more important than the remote future. In estimating the future of a fluctuation one can only make use of the experience of the past, so that it would seem more relevant to apply Dr. Good's remarks in relation to the past, rather than the future. In this case, however, we find the criterion needs to be weighted with t and not $1/(t+K)$.

I think it is more clear if one recalls that the criterion is concerned with errors in performance. The case of weighting function t requires that prediction of a transient (which is one of a set, known a priori) should be achieved with the minimum error as soon as possible.

Mr. J.F. Atherton questions my use of Taylor's series. A Taylor's series can certainly be employed to describe a continuous waveform over a finite region and provided one does not attempt to draw conclusions about the nature of the curve outside the given region this is certainly a permissible operation. This I have done, in effect, in the first part of the paper. In the second part I have endeavoured to deal with the discontinuous element for which Taylor's series cannot be employed.

Mr. D.J. Mynall and Mr. R.H. Barker both deal with interesting cases of usable unstable systems. Some facinating prospects are certainly brought to mind by the possibility of employing such systems. I have not admitted this possibility in the paper since I am concerned with "the history of waveforms upon which linear operations are performed" in which it is implied that such operations should be continuous. Both of the cases raised in discussion are usable, because they are in some sense discontinuous; in the one case by periodically "quenching" and in the other by discontinuously sampling.

Mr. H.C. Calpine makes an admirable contribution to the handling of the prediction problem when the signals may be described as quasi-stationary: that is stationary over lengthy periods but taking on different characteristics, over a known range, from time to time.

A word of warning might be appropriate concerning the difficulties of "synthesising" a required frequency spectrum from component parts. Unfortunately a synthesised spectrum, while differing only slightly from the frequency specification, can vary dramatically from its time specification, or impulse response which is what ultimately matters.

A further point in connection with the use of a "synthesised" predictor for quasi-stationary series is the necessity to disconnect the channels not required for a particular stationary section of the series. Unless this is done (by some biassing process, for example) no advantage will be obtained over a properly designed "optimum filter.

FIG. 1.

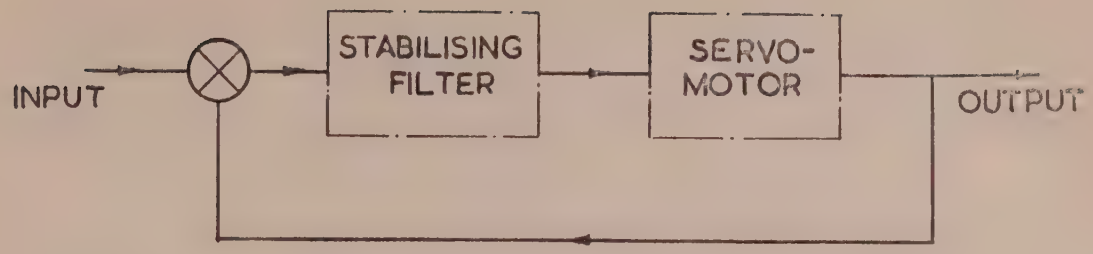


FIG. 2.

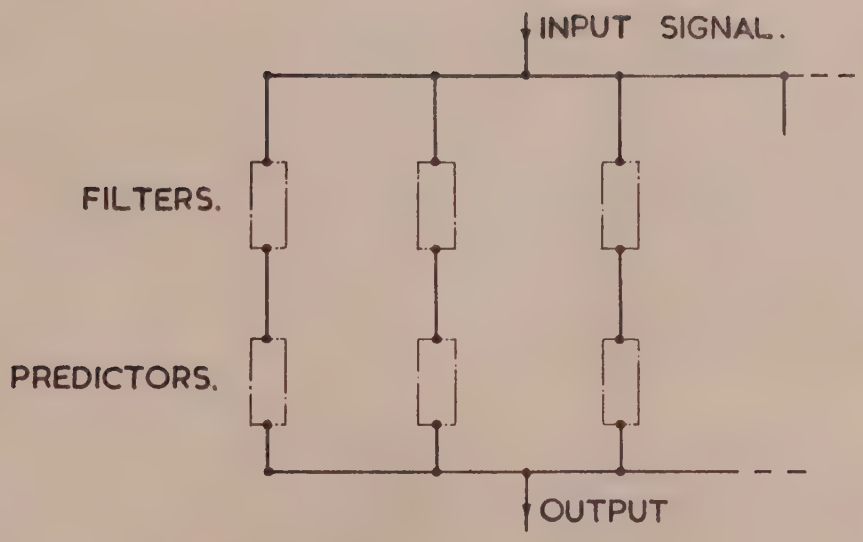
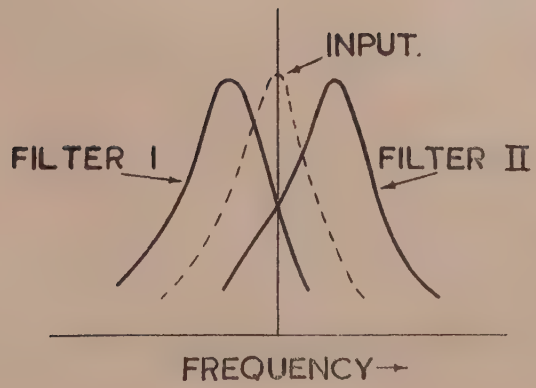
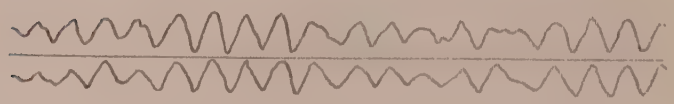


FIG. 3.



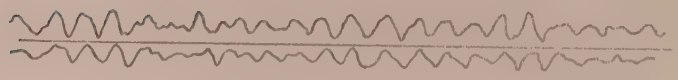
OPTIMUM PREDICTION.

NOISE FROM TUNED CIRCUIT.

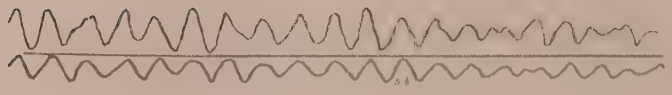


ORIGINAL.

0·2 PERIOD.



0·45 PERIOD.



ORIGINAL.

0·9 PERIOD

PREDICTION.

COMPARISON = OPTIMUM AND SYNTHETIC PREDICTION.

0·2 PERIOD.

0·45 PERIOD.

0·9 PERIOD.

EFFECT OF FILTERS.

DISCUSSION ON MR. MACKAY'S PAPER "ENTROPY, TIME AND INFORMATION"

PROF. M. S. BARTLETT

I apologise in advance for any nonsense my impromptu remarks might contain, but suggest that the concepts of entropy, time and information should be kept on a scientific level and as separate as possible from subjective notions. It seems to me that one quantitative idea emerging from the discussion so far about the passage of information from one system to another (compare the acquiring of "negative entropy" by living organisms at the expense of their environment) was a kind of conservation of information on entropy for the total system.

With regard to time; local systems usually appeal to the external environment to provide an independent time, so that time can hardly be said to be redundant except for a closed system.

A dichotomy into two systems e.g. observer and observed system, is essential to the notion of information begin communicated. For the total system, time might not only become redundant but would apparently (in the directional sense) cease to exist.

DR. I. J. GOOD

The analysis of the subjective passage of time by means of the amount of information passing into awareness seems at first sight to lead to difficulties. On the average the amount of information passing out of conscious awareness must be equal to that passing in. This shows that if Mr. MacKay is right it may be better to think in terms of the total amount of information in the whole mind, including what is unconscious.

The theory then has a verifiable consequence. When we grow older we find it more difficult to learn: we acquire information at a slower rate. The theory then implies that time should appear to pass faster. This is consistent with what elderly people usually assert. I do not, however, regard this argument as a very strong factor in favour of the theory.

PROF. J. L. VAN SOEST

Many years ago I have given in a biological conference a (possible) version on life in a thermodynamical sense, as a "fight against increasing entropy. &That is why the tree is growing up. Of course this fight is hopeless as seen as a problem of the living creature in his environment. Now, at the moment, information comes with good harmony into the problem: every living creature with a mind tries to increase its information-content.

&Afterwards W. J. Burgers raised an analogous question in a paper for the Kon. Ak. v. Wetenschappen, Amsterdam.

MR. W. S. PERCIVAL

I want to speak about 'time's arrow'. It is usually stated that time has, in effect, an arrow attached whereas the spatial coordinates have not. This is correct if we adopt Cartesian coordinates for space. However if we adopt spherical coordinates for space it is perfectly natural to endow their coordinate with an arrow pointing outwards. Such an arrow would be associated with divergence in a physical system.

The question therefore arises as to whether we can associate this arrow of space with the arrow of time. In other words can we associate the arrow of time with divergence in a physical system.

First let us consider an extreme case i.e. that of a flash of light. Now except in very special and artificial cases a flash of light always diverges and never converges. It is associated with an arrow pointing outwards in space.

But it is also quite clear that we must associate the term 'later' with a point on a wave front of greater radius, and the term 'earlier' with a point on a wave front of smaller radius. Hence in this case the arrow of space is directly associated with the arrow of time.

Now let us consider the case of a gas which is not in equilibrium e.g. it is crowded up into one end of the containing vessel. Such a gas will expand until it fills the vessel and is of substantially uniform density. Again we find the arrow of time associate with divergence i.e. with the outward pointing arrow of space.

Next let us consider a gas of uniform density, but of non-uniform temperature, so that the gas at one end of the containing figure is hot and at the other end is cold. Then both the hot and the cold portions will expand until the whole gas becomes of uniform temperature. Once again divergence is associated with what we call later in time. However it is not the divergence of the gas as a whole but of the non-uniformity of the gas. This is an important distinction to which we shall return in a moment.

Finally consider a gas in thermodynamic equilibrium. There is no divergence i.e. there is no motion which can be associated with the positive direction of the arrow of space. The motion that exists can be considered in a general sense as circulation but there is no average divergence. In this case we find that time's arrow also ceases to exist, and a succession of specifications of the system would exhibit no distinction between past and future.

What has this to do with Information Theory? Well, Mackay has suggested that a system should be considered later in time when it has given out more information. The giving out of information must physically be associated with a process of divergence which is seen in its clearest form in the divergence of a flash of light in which the distinction between earlier and later appears in its sharpest form.

Let us return to the special case of the gas of which one part is hot and the other cold. We have seen that such a gas will exhibit divergence, but the divergence, not of the gas as a whole, but of the non-uniformity of the gas. Now non-uniformity implies information and the divergence of the non-uniformity implies the passing outwards of information. Hence the outward pointing arrow of space which is associated with the arrow of time points in the direction of the outward flow of information.

I would suggest that this association of the arrow of time with the arrow of space is complementary to the ideas advanced by MacKay.

MR. D. M. MACKAY (In reply):

Professor Bartlett's remarks on the necessary dichotomy into observer and observed system are of much interest in relation to Whitehead's theory of abstractive hierarchies. In terms of this terminology, it would appear that time can enter only at the first level of abstraction and not at the level of the concrete. The metaphysical implications here might be profitably followed up.

I would agree with him that the conservation of information for a total system appears to be the working hypothesis on which we base our deductions (see Ref. 6). The law of conservation of energy may be a manifestation of the structural aspect of this same hypothesis.

With regard to the place of subjective notions, I would only insist that our scientific concepts ultimately find meaning in terms of our private sensations, and that to postpone a reckoning in these terms may be to court bankruptcy. My own view is that our flight from the subjective has carried us too far, and that our understanding of scientific ideas will be increased by a properly-disciplined restoration of perspective.

With Dr. Good's remarks on the non-linearity of the subjective timescale I am in agreement. The effect has in fact a possible objective parallel in the application of the theory to cosmology, in which the total information-content of the system again may provide the natural timescale. If the judgments of older people are admissible as evidence, however, I should be inclined to attribute their sensation to the fact that the fractional rate of increase in total awareness diminishes with age, so that even if information were steadily accumulated, "time" would appear to go more slowly. The evidence for this is equally to be found in young children. It is of course the total of awareness, conscious or unconscious, which is relevant, so long as it is included in the subconscious estimate to which we owe our subjective timescale.

Mr. Percival's spherical co-ordinate system has the disadvantage that it can only centre on one member of an assembly. Two sources of light could thus give oppositely-directed arrows in space, unless each had its own co-ordinate system. Indeed to adopt spherical co-ordinates is merely to correlate artificially the co-ordinate of time in one system with one dimension of space, so that the verification becomes a truism.

S.M.23863/R. O.S. (P. & S.).